



**THE  
INTERNATIONAL SERIES  
OF  
MONOGRAPHS ON PHYSICS**

**GENERAL EDITORS**

**†R. H. FOWLER, P. KAPITZA  
N. F. MOTT, E. C. BULLARD**

# THE INTERNATIONAL SERIES OF MONOGRAPHS ON PHYSICS

GENERAL EDITORS

THE LATE SIR RALPH FOWLER

P. KAPITZA

N. F. MOTT

E. C. BULLARD

Melville Wills Professor of Theoretical  
Physics in the University of Bristol.

Professor of Physics, University of  
Toronto.

## *Already Published*

- THE THEORY OF ELECTRIC AND MAGNETIC SUSCEPTIBILITIES. By J. H. VAN VLECK. 1932. Royal 8vo, pp. 396.
- THE THEORY OF ATOMIC COLLISIONS. By N. F. MOTT and H. S. W. MASSEY. 1933. Royal 8vo, pp. 300.
- RELATIVITY, THERMODYNAMICS, AND COSMOLOGY. By R. C. TOLMAN. 1934. Royal 8vo, pp. 518.
- ELECTROLYTES. By HANS FALKENHAGEN. Translated by R. P. BELL. 1934. Royal 8vo, pp. 364.
- CHEMICAL KINETICS AND CHAIN REACTIONS. By N. SEMENOFF. 1935. Royal 8vo, pp. 492.
- RELATIVITY, GRAVITATION, AND WORLD-STRUCTURE. By E. A. MILNE. 1935. Royal 8vo, pp. 378.
- THE QUANTUM THEORY OF RADIATION. By W. HEITLER. *Second Edition*. 1944. Royal 8vo, pp. 264.
- THEORETICAL ASTROPHYSICS: ATOMIC THEORY AND THE ANALYSIS OF STELLAR ATMOSPHERES AND ENVELOPES. By S. ROSSELAND. 1936. Royal 8vo, pp. 376.
- THE THEORY OF THE PROPERTIES OF METALS AND ALLOYS. By N. F. MOTT and H. JONES. 1936. Royal 8vo, pp. 340.
- ECLIPSES OF THE SUN AND MOON. By SIR FRANK DYSON and R. V. D. R. WOOLLEY. 1937. Royal 8vo, pp. 168.
- THE PRINCIPLES OF STATISTICAL MECHANICS. By R. C. TOLMAN. 1938. Royal 8vo, pp. 682.
- THE ULTRACENTRIFUGE. By THE SVEDBERG and KAI O. PEDERSEN. 1940. Royal 8vo, pp. 488.
- ELECTRONIC PROCESSES IN IONIC CRYSTALS. By N. F. MOTT and E. W. GURNEY. 1940. Royal 8vo, pp. 275.
- GEOMAGNETISM. By S. CHAPMAN and J. BARTELS. 1940. Royal 8vo, 2 vols., pp. 1076.
- THE SEPARATION OF GASES. By M. RUHEMANN. *Second Impression*. 1945. Royal 8vo, pp. 298.
- KINETIC THEORY OF LIQUIDS. By J. FRENKEL. 1946. Royal 8vo, pp. 500.
- THE PRINCIPLES OF QUANTUM MECHANICS. By P. A. M. DIRAC. *Third Edition*. 1947. Royal 8vo, pp. 324.
- COSMIC RAYS. By L. JÁNOSY. 1948. Royal 8vo, pp. 440.

# THEORY OF PROBABILITY

BY

HAROLD JEFFREYS

M.A., D.Sc., F.R.S.

PLUMIAN PROFESSOR OF ASTRONOMY  
UNIVERSITY OF CAMBRIDGE

*SECOND EDITION*

OXFORD  
AT THE CLARENDON PRESS

1948



*Oxford University Press, Amen House, London E.C. 4*

GLASGOW NEW YORK TORONTO MELBOURNE WELLINGTON

BOMBAY CALCUTTA MADRAS CAPE TOWN

*Geoffrey Cumberlege, Publisher to the University*

PRINTED IN GREAT BRITAIN

## PREFACE TO THE SECOND EDITION

IN the circumstances that have prevailed in the world since the appearance of this book, it is a welcome indication of increasing interest in the principles of scientific method that a second edition has been required. I have taken the opportunity to add some arguments that go far towards establishing the consistency of the product rule and therefore of the principle of inverse probability. A theory of invariance has been developed and applied to problems of estimation and significance, thus establishing the possibility of a consistent rule for stating prior probabilities over large parts of the subject. I am not satisfied that it is the only such rule or even the best one, but think that enough progress has been made to indicate that the attempt is worth pursuing.

I have not attempted to answer explicitly the criticisms made by reviewers, because on examination I found that they were all dealt with in the book already. What does strike me as remarkable is that no mention was made of the fact that the book contained useful methods of treatment of several problems of practical importance. I have still not gathered what distinction those statisticians who do not accept the epistemological approach draw between estimation problems and significance tests, or whether they think that they are saying anything about a hypothesis when they reject it. So far as I can judge from their pronouncements, they provide themselves with no reason against continuing to make predictions from it.

Several recent writers, especially in the United States, have described me as a follower of the late Lord Keynes. Without wishing to disparage Keynes, I must point out that the first two papers by Wrinch and me in the *Philosophical Magazine* of 1919 and 1921 preceded the publication of Keynes's book. What resemblance there is between the present theory and that of Keynes is due to the fact that Broad, Keynes, and my collaborator had all attended the lectures of W. E. Johnson. Keynes's distinctive contribution was the assumption that probabilities are only partially ordered; this contradicts my Axiom 1. I gave reasons for not accepting it in *Scientific Inference*. Keynes himself withdrew it in his biographical essay on F. P. Ramsey.

I have to thank several correspondents for suggesting corrections, especially Dr. H. Chojnacki-Hanani. Mr. P. H. Diananda of Caius College and Mr. V. S. Huzurbazar of Fitzwilliam House, Cambridge, have helped greatly in the proof-correction.

H. J.

ST. JOHN'S COLLEGE, CAMBRIDGE

October 1947

## PREFACE TO THE FIRST EDITION

THE chief object of this work is to provide a method of drawing inferences from observational data that will be self-consistent and can also be used in practice. Scientific method has grown up without much attention to logical foundations, and at present there is little relation between three main groups of workers. Philosophers, mainly interested in logical principles but not much concerned with specific applications, have mostly followed in the tradition of Bayes and Laplace; but with the brilliant exception of Professor C. D. Broad have not paid much attention to the consequences of adhering to the tradition in detail. Modern statisticians have developed extensive mathematical techniques, but for the most part have rejected the notion of the probability of a hypothesis, and thereby deprived themselves of any way of saying precisely what they mean when they decide between hypotheses. Physicists have been described, by an experimental physicist who has devoted much attention to the matter, as not only indifferent to fundamental analysis but actively hostile to it; and with few exceptions their statistical technique has hardly advanced beyond that of Laplace. In opposition to the statistical school, they and some other scientists are liable to say that a hypothesis is definitely proved by observation, which is certainly a logical fallacy; most statisticians appear to regard observations as a basis for possibly rejecting hypotheses, but in no case for supporting them. The latter attitude, if adopted consistently, would reduce all inductive inference to guesswork; the former, if adopted consistently, would make it impossible ever to alter the hypotheses, however badly they agreed with new evidence. The present attitudes of most physicists and statisticians are diametrically opposed, but lack of a common meeting-ground has, to a very large extent, prevented the opposition from being noticed. Nevertheless, both schools have made great scientific advances, in spite of the fact that their fundamental notions, for one reason or the other, would make such advances impossible if they were consistently maintained.

In the present book I reject the attempt to reduce induction to deduction, which is characteristic of both schools, and maintain that the ordinary common-sense notion of probability is capable of precise and consistent treatment when once an adequate language is provided for it. It leads to the result that a precisely stated hypothesis may attain either a high or a negligible probability as a result of observational data, and therefore to an attitude intermediate between those current in physics and statistics, but in accordance with ordinary

thought. Fundamentally the attitude is that of Bayes and Laplace, though it is found necessary to modify their hypotheses before some types of cases not considered by them can be treated, and some steps in the argument have been filled in. For instance, the rule for assessing probabilities given in the first few lines of Laplace's book is Theorem 7, and the principle of inverse probability is Theorem 10. There is, on the whole, a very good agreement with the recommendations made in statistical practice; my objection to current statistical theory is not so much to the way it is used as to the fact that it limits its scope at the outset in such a way that it cannot state the questions asked, or the answers to them, within the language that it provides for itself, and must either appeal to a feature of ordinary language that it has declared to be meaningless, or else produce arguments within its own language that will not bear inspection.

The most beneficial result that I can hope for as a consequence of this work is that more attention will be paid to the precise statement of the alternatives involved in the questions asked. It is sometimes considered a paradox that the answer depends not only on the observations but on the question; it should be a platitude.

The theory is applied to most of the main problems of statistics, and a number of specific applications are given. It is a necessary condition for their inclusion that they shall have interested me. As my object is to produce a general method I have taken examples from a number of subjects, though naturally there are more from physics than from biology and more from geophysics than from atomic physics. It was, as a matter of fact, mostly with a view to geophysical applications that the theory was developed. It is not easy, however, to produce a statistical method that has application to only one subject; though intraclass correlation, for instance, which is a matter of valuable positive discovery in biology, is usually an unmitigated nuisance in physics. It may be felt that many of the applications suggest further questions. That is inevitable. It is usually only when one group of questions has been answered that a further group can be stated in an answerable form at all.

I must offer my warmest thanks to Professor R. A. Fisher and Dr. J. Wishart for their kindness in answering numerous questions from a not very docile pupil, and to Mr. R. B. Braithwaite, who looked over the manuscript and suggested a number of improvements; also to the Clarendon Press for their extreme courtesy at all stages.

H. J.

ST. JOHN'S COLLEGE, CAMBRIDGE

## CONTENTS

I. FUNDAMENTAL NOTIONS	1
II. DIRECT PROBABILITIES	47
III. ESTIMATION PROBLEMS	99
IV. APPROXIMATE METHODS AND SIMPLIFICATIONS	168
V. SIGNIFICANCE TESTS: ONE NEW PARAMETER	220
VI. SIGNIFICANCE TESTS: VARIOUS COMPLICATIONS	305
VII. FREQUENCY DEFINITIONS AND DIRECT METHODS	341
VIII. GENERAL QUESTIONS	372
APPENDIX. TABLES OF $K$	396
NOTE ON THE CONSISTENCY OF THE PRODUCT RULE	405
NOTE ON THE INFINITE REGRESS ARGUMENT	407
INDEX	408

## I

### FUNDAMENTAL NOTIONS

They say that Understanding ought to work by the rules of right reason. These rules are, or ought to be, contained in Logic; but the actual science of logic is conversant at present only with things either certain, impossible, or entirely doubtful, none of which (fortunately) we have to reason on. Therefore the true logic for this world is the calculus of Probabilities, which takes account of the magnitude of the probability which is, or ought to be, in a reasonable man's mind.

J. CLERK MAXWELL

1.0. THE fundamental problem of scientific progress, and a fundamental one of everyday life, is that of learning from experience. Knowledge obtained in this way is partly merely description of what we have already observed, but part consists of making inferences from past experience to predict future experience. This part may be called generalization or induction. It is the most important part; events that are merely described and have no apparent relation to others may as well be forgotten, and in fact usually are. The theory of learning in general is the branch of logic known as epistemology. A few illustrations will indicate the scope of induction. A botanist is confident that the plant that grows from a mustard seed will have yellow flowers with four long and two short stamens, and four petals and sepals, and this is inferred from previous instances. The *Nautical Almanac's* predictions of the positions of the planets, an engineer's estimate of the output of a new dynamo, and an agricultural statistician's advice to a farmer about the utility of a fertilizer are all inferences from past experience. When a musical composer scores a bar he is expecting a definite series of sounds when an orchestra carries out his instructions. In every case the inference rests on past experience that certain relations have been found to hold; and those relations are then applied to new cases that were not part of the original data. The same applies to my expectations about the flavour of my next meal. The process is so habitual that we hardly notice it, and we can hardly exist for a minute without carrying it out. On the rare occasions when anybody mentions it, it is called common sense and left at that.

Now such inference is not covered by logic, as the word is ordinarily understood. Traditional or deductive logic admits only three attitudes to any proposition: definite proof, disproof, or blank ignorance. But no number of previous instances of a rule will provide a deductive proof

that the rule will hold in a new instance. There is always the formal possibility of an exception.

Deductive logic and its close associate, pure mathematics, have been developed to an enormous extent, and in a thoroughly systematic way—indeed several ways. Scientific method, on the other hand, has grown up more or less haphazard, techniques being developed to deal with problems as they arose, without much attempt to unify them, except so far as most of the theoretical side involved the use of pure mathematics, the teaching of which required attention to the nature of some sort of proof. Unfortunately the mathematical proof is deductive, and induction in the scientific sense is simply unintelligible to the pure mathematician—as such; in his unofficial capacity he may be able to do it very well. Consequently little attention has been paid to the nature of induction, and apart from actual mathematical technique the relation between science and mathematics has done little to develop a connected account of the characteristic scientific mode of reasoning. Many works exist claiming to give such an account, and there are some highly useful ones dealing with methods of treating observations that have been found useful in the past and may be found useful again. But when they try to deal with the underlying general theory they suffer from all the faults that modern pure mathematics has been trying to get rid of: self-contradictions, circular arguments, postulates used without being stated, and postulates stated without being used. Running through the whole is the tendency to claim that scientific method can be reduced in some way to deductive logic, which is the most fundamental fallacy of all: it can be done only by rejecting its chief feature, induction.

The principal field of application of deductive logic is pure mathematics, which pure mathematicians recognize quite frankly as dealing with the working out of the consequences of stated rules with no reference to whether there is anything in the world that satisfies those rules. Its propositions are of the form ‘If  $p$  is true, then  $q$  is true’, irrespective of whether we can find any actual instance where  $p$  is true. The mathematical proposition is the *whole* proposition, ‘If  $p$  is true, then  $q$  is true’, which may be true even if  $p$  is in fact always false. In applied mathematics, as usually taught, general rules are asserted as applicable to the external world, and the consequences are developed logically by the technique of pure mathematics. If we inquire what reason there is to suppose the general rules true, the usual answer is simply that they are known from experience. However, this use of the

word 'experience' covers a confusion. The rules are inferred from past experience, and then applied to future experience, which is not the same thing. There is no guarantee whatever in deductive logic that a rule that has held in all previous instances will not break down in the next instance or in all future instances. Indeed there are an infinite number of rules that have held in all previous cases and cannot possibly all hold in future ones. For instance, consider a body falling freely under gravity. It would be asserted that the distance at time  $t$  below a fixed level is given by a formula of the type

$$s = a + ut + \frac{1}{2}gt^2. \quad (1)$$

This might be asserted from observations of  $s$  at a series of instants  $t_1, t_2, \dots, t_n$ . That is, our previous experience asserts the proposition that  $a, u$ , and  $g$  exist such that

$$s_r = a + ut_r + \frac{1}{2}gt_r^2 \quad (2)$$

for all values of  $r$  from 1 to  $n$ . But the law (1) is asserted for *all* values of  $t$ . But consider the law

$$s = a + ut + \frac{1}{2}gt^2 + f(t)(t-t_1)(t-t_2)\dots(t-t_n), \quad (3)$$

where  $f(t)$  may be any function whatever that is not infinite at any of  $t_1, t_2, \dots, t_n$ , and  $a, u$ , and  $g$  have the same values as in (1). There are an infinite number of such functions. Every form of (3) will satisfy the set of relations (2), and therefore every one has held in all previous cases. But if we consider any other instant  $t_{n+1}$  (which might be either within or outside the range of time between the first and last of the original observations) it will be possible to choose  $f(t_{n+1})$  in such a way as to give  $s$  as found from (3) any value whatever at time  $t_{n+1}$ . Further, there will be an infinite number of forms of  $f(t)$  that would give the same value of  $f(t_{n+1})$ , and there are an infinite number that would give different values. If we observe  $s$  at time  $t_{n+1}$ , we can choose  $f(t_{n+1})$  to give agreement with it, but an infinite number of forms of  $f(t)$  consistent with this value would be consistent with any arbitrary value of  $s$  at a further moment  $t_{n+2}$ . That is, even if all the observed values agree with (1) exactly, deductive logic can say nothing whatever about the value of  $s$  at any other time. An infinite number of laws agree with previous experience, and an infinite number that have agreed with previous experience will inevitably be wrong in the next instance. What the applied mathematician does, in fact, is to select one form out of this infinity; and his reason for doing so has nothing whatever to do with traditional logic. He chooses the simplest. This is actually an understatement of the case; because in general the observations will not agree with (1)



exactly, a polynomial of  $n$  terms can still be found that will agree exactly with the observed values at times  $t_1, \dots, t_n$ , and yet the form (1) may still be asserted. Similar considerations apply to any quantitative law. The further discussion of this matter must be reserved till we come to significance tests. We need notice at the moment only that the choice of the simplest law that fits the facts is an essential part of procedure in applied mathematics, and cannot be justified by the methods of deductive logic. It is, however, rarely stated, and when it is stated it is usually in a manner suggesting that it is something to be ashamed of. We may recall the words of Brutus.

But 'tis a common proof  
That lowliness is young ambition's ladder,  
Whereto the climber upwards turns his face;  
But when he once attains the upmost round,  
He then unto the ladder turns his back,  
Looks in the clouds, scorning the base degrees  
By which he did ascend.

It is asserted, for instance, that the choice of the simplest law is purely a matter of economy of description or thought, and has nothing to do with any reason for believing the law. No reason in deductive logic, certainly; but the question is, Does deductive logic contain the whole of reason? It does give economy of description of past experience, but is it unreasonable to be interested in future experience? Do we make predictions merely because those predictions are the easiest to make? Does the Nautical Almanac Office laboriously work out the positions of the planets by means of a complicated set of tables based on the law of gravitation and previous observations, merely for convenience, when it might much more easily guess them? Do sailors trust the safety of their ships to the accuracy of these predictions for the same reason? Does a town install a new tramway system, with expensive plant and much preliminary consultation with engineers, with no more reason to suppose that the trams will move than that the laws of electromagnetic induction are a saving of trouble? I do not believe for a moment that anybody will answer any of these questions in the affirmative; but an affirmative answer is implied by the assertion that is still frequently made, that the choice of the simplest law is merely a matter of convention. I say, on the contrary, that the simplest law is chosen because it is the most likely to give correct predictions; that the choice is based on a reasonable degree of belief; and that the fact that deductive logic provides no explanation of the choice of the simplest law is an absolute proof that deductive logic is grossly inadequate to

cover scientific and practical requirements. It is sometimes said, again, that the trust in the simple law is a peculiarity of human psychology; a different type of being might behave differently. Well, I see no point whatever in discussing at length whether the human mind is any use; it is not a perfect reasoning instrument, but it is the only one we have. Deductive logic itself could never be known without the human mind. If anybody rejects the human mind and then holds that he is constructing valid arguments, he is contradicting himself; if he holds that human minds other than his own are useless, and then hopes to convince them by argument, he is again contradicting himself. A critic is himself using inductive inference when he expects his words to convey the same meaning to his audience as they do to himself, since the meanings of words are learned first by noting the correspondence between things and the sounds uttered by other people, and then applied in new instances. On the face of it, it would appear that a general statement that something accepted by the bulk of mankind is intrinsically nonsense requires much more to support it than a mere declaration.

Many attempts have been made, while accepting induction, to claim that it can be reduced in some way to deduction. Bertrand Russell has remarked that induction is either disguised deduction or a mere method of making plausible guesses.† In the former sense we must look for some general principle, which states a set of possible alternatives; then observations are used to show that all but one of these are wrong, and the survivor is held to be deductively demonstrated. Such an attitude has been widely advocated. On it I quote Professor C. D. Broad.‡

‘The usual view of the logic books seems to be that inductive arguments are really syllogisms with propositions summing up the relevant observations as minors, and a common major consisting of some universal proposition about nature. If this were true it ought to be easy enough to find the missing major, and the singular obscurity in which it is enshrouded would be quite inexplicable. It is reverently referred to by inductive logicians as the Uniformity of Nature; but, as it is either never stated at all or stated in such terms that it could not possibly do what is required of it, it appears to be the inductive equivalent of Mrs. Gamp’s mysterious friend, and might be more appropriately named Major Harris.

† *Principles of Mathematics*, p. 380. He said, at the Aristotelian Society summer meeting in 1938, that this remark has been too much quoted. I therefore offer apologies for quoting it again. He has also remarked that the inductive philosophers of Central Africa formerly held the view that all men were black. My comment would be that the deductive ones, if there were any, did not hold that there were any men, black, white, or yellow.

‡ *Mind*, 29, 1920, 11.

'It is in fact easy to prove that this whole way of looking at inductive arguments is mistaken. On this view they are all syllogisms with a common major. Now their minors are propositions summing up the relevant observations. If the observations have been carefully made the minors are practically certain. Hence, if this theory were true, the conclusions of all inductive arguments in which the observations were equally carefully made would be equally probable. For what could vary the probabilities? Not the major, which is common to all of them. Not the minors, which by hypothesis are equally certain. Not the mode of reasoning, which is syllogistic in each case. But the result is preposterous, and is enough to refute the theory which leads to it.'

Attempts have been made recently to supply the missing major by several modern physicists, notably Sir Arthur Eddington and Professor E. A. Milne. But their general principles and their results differ even within the very limited field of knowledge where they have been applied. How is a person with less penetration to know which is right, if any? Only by comparing the results with observation; and then his reason for believing the survivor to be likely to give the right results in future is inductive. I am not denying that one of them may have got the right results. But I reject the statement that any of them can be said to be certainly right as a matter of pure logic, independently of experience; and I gravely doubt whether any of them could have been thought of at all had the authors been unaware of the vast amount of previous work that had led to the establishment by inductive methods of the laws that they set out to explain. These attempts, though they appear to avoid Broad's objection, do so only within a limited range, and it is doubtful whether such an attempt is worth making if it can at best achieve a partial success, when induction can cover the whole field without supposing that special rules hold in certain subjects.

I should maintain (with N. R. Campbell, who says† that a physicist would be more likely to interchange the two terms in Russell's statement) that a great deal of what passes for deduction is really disguised induction, and that even some of the postulates of *Principia Mathematica* are adopted on inductive grounds (which, incidentally, are false).

Two attempts at a justification of induction, still sometimes made, are as follows. (1) Induction has worked in the past; therefore it will work in the future. It is obvious that this is itself an inductive inference and involves the same problems in a more complicated way. (2) The struggle for existence would favour members with the ability to predict correctly the consequences of their actions. Consequently the fact that man has survived implies that he has this ability (and presumably

† *Physics, The Elements*, 1920, 9.

*Amoeba* has too). But the belief that there is a struggle for existence and that it favours particular types is based on induction. Both arguments replace the original question by another as difficult or more so, and take no effective step towards a solution.

Karl Pearson† writes as follows:

'Now this is the peculiarity of scientific method, that when once it has become a habit of mind, that mind converts all facts whatsoever into science. The field of science is unlimited; its material is endless, every group of natural phenomena, every phase of social life, every stage of past or present development is material for science. *The unity of all science consists alone in its method, not in its material.* The man who classifies facts of any kind whatever, who sees their mutual relation and describes their sequences, is applying the scientific method and is a man of science. The facts may belong to the past history of mankind, to the social statistics of our great cities, to the atmosphere of the most distant stars, to the digestive organs of a worm, or to the life of a scarcely visible bacillus. It is not the facts themselves which form science, but the methods by which they are dealt with.'

Here, in a few sentences, Pearson sets our problem. The italics are his. He makes a clear distinction between method and material. No matter what the subject-matter, the fundamental principles of the method must be the same. There must be a uniform standard of validity for all hypotheses, irrespective of the subject. Different laws may hold in different subjects, but they must be tested by the same criteria; otherwise we have no guarantee that our decisions will be those warranted by the data and not merely the result of inadequate analysis or of believing what we want to believe. An adequate theory of induction must satisfy two conditions. First, it must provide a general method; secondly, the principles of the method must not of themselves say anything about the world. If the rules are not general, we shall have different standards of validity in different subjects, or different standards for one's own hypotheses and somebody else's. If the rules of themselves say anything about the world, they will make empirical statements independently of observational evidence, and thereby limit the scope of what we can find out by observation. If there are such limits, they must be inferred from observation; we must not assert them in advance.

We must notice at the outset that induction is more general than deduction. The answers given by the latter are limited to a simple 'yes', 'no', or 'it doesn't follow'. Inductive logic must split up the last alternative, which is of no interest to deductive logic, into a number of others, and say which of them it is most reasonable to believe on

† *The Grammar of Science*, 1892. P. 16 of Everyman edition, 1938.

the evidence available. Complete proof and disproof are merely the extreme cases. Any inductive inference involves in its very nature the possibility that the alternative chosen as the most likely may in fact be wrong. Exceptions are always possible, and if a theory does not provide for them it will be claiming to be deductive when it cannot be. On account of this extra generality, induction must involve postulates not included in deduction. Our problem is to state these postulates. It is important to notice that they cannot be proved by deductive logic. If they could, induction would be reduced to deduction, which is impossible. Equally they are not empirical generalizations; for induction would be needed to make them and the argument would be circular. We must in fact distinguish the general rules of the theory from the empirical content. The general rules are *a priori* propositions, accepted independently of experience, and making by themselves no statement about experience. Induction is the application of the rules to observational data.

Our object, in short, is not to prove induction; it is to tidy it up. Even among professional statisticians there are considerable differences about the best way of treating the same problem, and, I think, all statisticians would reject some methods habitual in some branches of physics. The question is whether we can construct a general method, the acceptance of which would avoid these differences or at least reduce them.

1.1. The test of the general rules, then, is not any sort of proof. This is no objection because the primitive propositions of deductive logic cannot be proved either. All that can be done is to state a set of hypotheses, as plausible as possible, and see where they lead us. The fullest development of deductive logic and of the foundations of mathematics is that of *Principia Mathematica*, which starts with a number of primitive propositions taken as axioms; if the conclusions are accepted, that is because we are willing to accept the axioms, not because the latter are proved. The same applies, or used to apply, to Euclid. We must not hope to prove our primitive propositions when this is the position in pure mathematics itself. But we have rules to guide us in stating them, largely suggested by the procedure of logicians and pure mathematicians.

1. All hypotheses used must be explicitly stated, and the conclusions must follow from the hypotheses.

2. The theory must be self-consistent; that is, it must not be possible

to derive contradictory conclusions from the postulates and any given set of observational data.

3. Any rule given must be applicable in practice. A definition is useless unless the thing defined can be recognized in terms of the definition when it occurs. The existence of a thing or the estimate of a quantity must not involve an impossible experiment.

4. The theory must provide explicitly for the possibility that inferences made by it may turn out to be wrong. A law may contain adjustable parameters, which may be wrongly estimated, or the law itself may be afterwards found to need modification. It is a fact that revision of scientific laws has often been found necessary in order to take account of new information—the relativity and quantum theories providing conspicuous instances—and there is no conclusive reason to suppose that any of our present laws are final. But we do accept inductive inference in some sense; we have a certain amount of confidence that it will be right in any particular case, though this confidence does not amount to logical certainty.

5. The theory must not deny any empirical proposition *a priori*; any precisely stated empirical proposition must be formally capable of being accepted, in the sense of the last rule, given a moderate amount of relevant evidence.

These five rules are essential. The first two impose on inductive logic criteria already required in pure mathematics. The third and fifth enforce the distinction between *a priori* and empirical propositions; if an existence depends on an inapplicable definition we must either find an applicable one, treat the existence as an empirical proposition requiring test, or abandon it. The fourth states the distinction between induction and deduction. The fifth makes Pearson's distinction between material and method explicit, and involves the definite rejection of attempts to derive empirically verifiable propositions from general principles adopted independently of experience.

The following rules also serve as useful guides.

6. The number of postulates should be reduced to a minimum. This is done for deductive logic in *Principia*, though many theorems proved there appear to be as obvious intuitively as the postulates. The motive for not accepting other obvious propositions as postulates is partly artistic. But we cannot regard the human mind as a perfect reasoner, and a reduction of the number of postulates affords a check on the consistency of different propositions, any of which we might be ready to accept by itself. This is still more needed in induction, since the

beliefs often accepted as intuitively certain are more numerous, and, I believe, some of them are definitely inconsistent, while others are not primitive propositions but inductive inferences. If they are, they cannot, of course, be asserted as certain, but they may be asserted with so high a probability that there will be little difference in practice.

7. While we do not regard the human mind as a perfect reasoner, we must accept it as a useful one and the only one available. The theory need not represent actual thought-processes in detail, but should agree with them in outline. We are not limited to considering only the thought-processes that people describe to us. It often happens that their behaviour is a better criterion of their inductive processes than their arguments. If a result is alleged to be obtained by arguments that are certainly wrong, it does not follow that the result is wrong, since it may have been obtained by a rough inductive process that the author thinks it undesirable or unnecessary to state on account of the traditional insistence on deduction as the only valid reasoning. I disagree utterly with many arguments produced by the chief current schools of statistics, but I rarely differ seriously from the conclusions; their practice is far better than their precept. I should say that this is the result of common sense emerging in spite of the deficiencies of mathematical teaching. The theory must provide criteria for testing the chief types of scientific law that have actually been suggested or asserted. Any such law must be taken seriously in the sense that it can be asserted with confidence on a moderate amount of evidence. The fact that simple laws are often asserted will, on this criterion, require us to say that in any particular instance some simple law is quite likely to be true.

8. In view of the greater complexity of induction, we cannot hope to develop it more thoroughly than deduction. We shall therefore take it as a rule that an objection carries no weight if an analogous objection would invalidate part of generally accepted pure mathematics. I do not wish to insist on any particular justification of pure mathematics, since authorities on its foundations are far from being agreed among themselves. In *Principia* much of higher mathematics, including the whole theory of the continuous variable, rests on the axioms of infinity and reducibility, which are rejected by Hilbert. F. P. Ramsey rejects the axiom of reducibility, while declaring that the multiplicative axiom, properly stated, is the most evident tautology, though Whitehead and Russell express much doubt about it and carefully separate propositions that depend on it from those that can be proved without it. I should

go further and say that the proof of the existence of numbers, according to the *Principia* definition of number, depends on the postulate that all individuals are permanent, which is an empirical proposition, and a false one, and should not be made part of a deductive logic. But we do not need such a proof for our purposes. It is enough that pure mathematics should be consistent. If the postulate could hold in *some* world, even if it was not the actual world, that would be enough to establish consistency. Then the derivation of ordinary mathematics from the postulates of *Principia* can be regarded as a proof of its consistency. But the justification of all the justifications seems to be that they lead to ordinary pure mathematics in the end; I shall assume that the latter has validity irrespective of any particular justification.

The above principles will strike many readers as platitudes; and if they do I shall not object. But they require the rejection of several principles accepted as fundamental in other theories. They rule out, in the first place, any definition of probability that attempts to define probability in terms of infinite sets of possible observations, for we cannot in practice make an infinite number of observations. The Venn limit, the hypothetical infinite population of Fisher, and the ensemble of Willard Gibbs are useless to us by rule 3. Though many accepted results appear to be based on these definitions, a closer analysis shows that further hypotheses are required before any results are obtained, and these hypotheses are not stated. In fact, no 'objective' definition of probability in terms of actual or possible observations, or possible properties of the world, is admissible. For, if we made anything in our fundamental principles depend on observations or on the structure of the world, we should have to say either (1) that the observations we can make, and the structure of the world, are initially unknown; then we cannot know our fundamental principles, and we have no possible starting-point; or (2) that we know *a priori* something about observations or the structure of the world, and this is illegitimate by rule 5. Attempts to use the latter principle will superpose our preconceived notions of what is objective on the entire system, whereas, if objectivity has any meaning at all, our aim must be to *find out* what is objective by means of observations. To try to give objective definitions at the start will at best produce a circular argument, may lead to contradictions, and in any case will make the whole scheme subjective beyond hope of recovery. We must not rule out any empirical proposition *a priori*; we must provide a system that will enable us to test it when occasion arises, and this requires a completely comprehensive formal scheme.



We must also reject what is variously called the principle of causality, determinism, or the uniformity of nature, in any such form as 'Precisely similar antecedents lead to precisely similar consequences'. No two sets of antecedents are ever identical; they must differ at least in time and position. But even if we decide to regard time and position as irrelevant (which may be true, but has no justification in pure logic) the antecedents are never identical. In fact, determinists usually recognize this verbally and try to save the principle by restating it in some such form as: 'In precisely the same circumstances very similar things can be observed, or very similar things can usually be observed.'<sup>†</sup> If 'precisely the same' is intended to be a matter of absolute truth, we cannot achieve it. Astronomy is usually considered a science, but the planets have never even approximately repeated their positions since astronomy began. The principle gives us no means of inferring the accelerations at a single instant, and is utterly useless. Further, if it was to be any use we should have to *know* at any application that the entire condition of the world was the same as in some previous instance. This is never satisfied in the most carefully controlled experimental conditions. The most that can be done is to make those conditions the same that we believe to be relevant—'the same' can never in practice mean more than 'the same as far as we know', and usually means a great deal less. The question then arises, How do we know that the neglected variables are irrelevant? Only by actually allowing them to vary and verifying that there is no associated variation in the result; but this requires the use of significance tests, a theory of which must therefore be given before there is any application of the principle, and when it is given it is found that the principle is no longer needed and can be omitted by rule 6. It may conceivably be true in some sense, though nobody has succeeded in stating clearly what this sense is. But what is quite certain is that it is useless.

Causality, as used in applied mathematics, has a more general form, such as: 'Physical laws are expressible by mathematical equations, possibly connecting continuous variables, such that in any case, given a finite number of parameters, some variable or set of variables that appears in the equations is uniquely determined in terms of the others.' This does not require that the values of the relevant parameters should be actually repeated; it is possible for an electrical engineer to predict the performance of a dynamo without there having already been some exactly similar dynamo. The equations, which we call laws, are inferred

<sup>†</sup> W. H. George, *The Scientist in Action*, 1936, p. 48.

from previous instances and then applied to instances where the relevant quantities are different. This form permits astronomical prediction. But it still leaves the questions 'How do we know that no other parameters than those stated are needed?', 'How do we know that we need consider no variables as relevant other than those mentioned explicitly in the laws?', and 'Why do we believe the laws themselves?' It is only after these questions have been answered that we can make any actual application of the principle, and the principle is useless until we have attended to the epistemological problems. Further, the principle happens to be false for quantitative observations. It is not true that observed results agree exactly with the predictions made by the laws actually used. The most that the laws do is to predict a variation that accounts for the greater part of the observed variation; it never accounts for the whole. The balance is called 'error' and usually quickly forgotten or altogether disregarded in physical writings, but its existence compels us to say that the laws of applied mathematics do not express the whole of the variation. Their justification cannot be exact mathematical agreement, but only a partial one depending on what fraction of the observed variation in one quantity is accounted for by the variations of the others. The phenomenon of error is often dealt with by a suggestion of various minor variations that might alter the measurements, but this is no answer. An exact quantitative prediction could never be made, even if such a suggestion was true, unless we knew in each individual case the actual amounts of the minor variations, and we never do. If we did we should allow for them and obtain a still closer agreement; but the fact remains that in practice, however fully we take small variations into account, we never get exact agreement. A physical law, for practical use, cannot be merely a statement of exact predictions; if it was it would invariably be wrong and would be rejected at the next trial. Quantitative prediction must always be prediction within a margin of uncertainty; the amount of this margin will be different in different cases, but for a law to be of any use it must state the margin explicitly. The outstanding variation, for practical application, is as essential a part of the law as the predicted variation is, and a valid statement of the law must express it. But in any individual case this outstanding variation is not known. We know only something about its possible range of values, not what the actual value will be. *Hence a physical law is not an exact prediction, but a statement of the relative probabilities of variations of different amounts. It is only in this form that we can avoid rejecting causality altogether as false,*

*or as inapplicable under rule 3; but a statement of ignorance of the individual errors has become an essential part of it, and we must recognize that the physical law itself, if it is to be of any use, must have an epistemological content.*

The impossibility of exact prediction has recently been forced on the attention of physicists by Heisenberg's Uncertainty Principle. It is remarkable, considering that the phenomenon of errors of observation was discussed by Laplace and Gauss, that there should still have been any physicists that thought that actual observations were exactly predictable; yet attempts to evade the principle have shown that many exist. The principle is actually no new uncertainty. What Heisenberg has done is to consider the most refined types of observation that modern physics suggests might be possible, and to obtain a lower limit to the uncertainty; but it is much smaller than the old uncertainty, which was never neglected except by misplaced optimism. The existence of errors of observation seems to have escaped the attention of many philosophers that have discussed the uncertainty principle; this is perhaps because they tend to get their notions of physics from popular writings, and not from works on the combination of observations. Their criticisms of popular physics, mostly valid as far as they go, would gain enormously in force if they attended to what we knew about errors before Heisenberg.†

The word *error* is liable to be interpreted in some ethical sense, but its scientific meaning is closely connected with the original one. Latin *errare*, in its original sense, means to wander, not to sin or to make a mistake. The meaning occurs in 'knight-errant'. The error means simply the outstanding variation after we have done our best to interpret the whole variation.

The criterion of universal assent, stated by Dr. N. R. Campbell and by Professor H. Dingle in his *Science and Human Experience* (but abandoned in his *Through Science to Philosophy*), must also be rejected

† Professor L. S. Stebbing (*Philosophy and the Physicists*, 1938, p. 198) remarks: 'There can be no doubt at all that precise predictions concerning the behaviour of macroscopic bodies are made and are *exactly* verified within the limits of experimental error.' Without the saving phrase at the end the statement is intelligible, and false. With it, it is meaningless. The severe criticism of much in modern physics contained in this book is, in my opinion, thoroughly justified, but the later parts lose much of their point through inattention to the problem of errors of observation. Some philosophers, however, have seen the point quite clearly. For instance, Professor J. H. Muirhead (*The Elements of Ethics*, 1910, pp. 37-8) states: 'The truth is that what is called a natural law is itself not so much a statement of fact as of a standard or type to which facts have been found more or less to approximate. This is true even in inorganic nature.' I am indebted to Mr. John Bradley for the reference.

by rule 3. This criterion requires general acceptance of a principle before it can be adopted. But it is impossible to ask everybody's consent before one believes anything; and if 'everybody' is replaced by 'everybody qualified to judge', we cannot apply the criterion until we know who is qualified, and even then it is liable to happen that only a small fraction of the people capable of expressing an opinion on a scientific paper read it at all, and few even of those do express any. Campbell lays much stress on a physicist's characteristic intuition,<sup>†</sup> which apparently enables him always to guess right. But if there is any such intuition there is no need for the criterion of general agreement or for any other. The need for some general criterion is that even among those apparently qualified to judge there are often serious differences of opinion about the proper interpretation of the same facts; what we need is an impersonal criterion that will enable an individual to see whether, in any particular instance, he is following the rules that other people follow and that he himself follows in other instances.

**1.2.** The chief constructive rule is 4. It declares that there is a valid primitive idea expressing the degree of confidence that we may reasonably have in a proposition, even though we may not be able to give either a deductive proof or a disproof of it. In extreme cases it may be a mere statement of ignorance. We need to express its rules. One obvious one (though it is very commonly overlooked) is that it depends both on the proposition considered and on the data in relation to which it is considered. Suppose that I know that Smith is an Englishman, but otherwise know nothing particular about him. He is very likely, on that evidence, to have a blue right eye. But suppose that I am informed that his left eye is brown—the probability is changed completely. This is a trivial case, but the principle in it constitutes most of our subject-matter. It is a fact that our degrees of confidence in a proposition habitually change when we make new observations or new evidence is communicated to us by somebody else, and this change constitutes the essential feature of all learning from experience. We must therefore be able to express it. Our fundamental idea will not be simply the probability of a proposition  $p$ , but the probability of  $p$  on data  $q$ . Omission to recognize that a probability is a function of two arguments, both propositions, is responsible for a large number of serious mistakes; in some hands it has led to correct results, but at the

<sup>†</sup> *Aristot. Soc. Suppl.* vol. 17, 1938, 122.

cost of omitting to state essential hypotheses and giving a delusive appearance of simplicity to what are really very difficult arguments. *It is no more valid to speak of the probability of a proposition without stating the data than it would be to speak of the value of  $x+y$  for given  $x$ , irrespective of the value of  $y$ .*

We can now proceed on rule 7. It is generally believed that probabilities are orderable: that is, that if  $p$ ,  $q$ ,  $r$  are three propositions, the statement 'on data  $p$ ,  $q$  is more probable than  $r$ ' has a meaning. In actual cases people may disagree about which is the more probable, and it is sometimes said that this implies that the statement has no meaning. But the differences may have other explanations: (1) The commonest is that the probabilities are on different data, one person having relevant information not available to the other, and we have made it an essential point that the probability depends on the data. The conclusion to draw in such a case is that, if people argue without telling each other what relevant information they have, they are wasting their time. (2) The estimates may be wrong. It is perfectly possible to get a wrong answer in pure mathematics, so that by rule 8 this is no objection. In this case, where the probability is often a mere guess, we cannot expect the answer to be right, though it may be and often is a fair approximation. (3) The wish may be father to the thought. But perhaps this also has an analogue in pure mathematics, if we consider the number of fallacious methods of squaring the circle and proving Fermat's last theorem that have been given, merely because people wanted  $\pi$  to be an algebraic or rational number or the theorem to be true. In any case alternative hypotheses are open to the same objection, on the one hand, that they depend on a wish to have a wholly deductive system and to avoid the explicit statement of the fact that scientific inferences are not certain; or, on the other, that the statement that there is a most probable alternative on given data may curtail their freedom to believe another when they find it more pleasant. I think that these reasons account for all the apparent differences, but they are not fundamental. Even if people disagree about which is the more probable alternative, they agree that the comparison has a meaning. We shall assume that this is right. The meaning, however, is not a statement about the external world; it is a relation of inductive logic. Our primitive notion, then, is that of the relation 'given  $p$ ,  $q$  is more probable than  $r$ ', where  $p$ ,  $q$ , and  $r$  are three propositions. If this is satisfied in a particular instance, we say that  $r$  is less probable than  $q$ , given  $p$ ; this is the definition of *less probable*. If given  $p$ ,  $q$  is neither

more nor less probable than  $r$ ,  $q$  and  $r$  are *equally probable*, given  $p$ . Then our first axiom is

AXIOM 1. *Given  $p$ ,  $q$  is either more, equally, or less probable than  $r$ , and no two of these alternatives can be true.*

This axiom may be called that of the comparability of probabilities. In *Scientific Inference* I took it in a more general form, assuming that the probabilities of propositions on different data can be compared. But this appears to be unnecessary, because it is found that the comparability of probabilities on different data, whenever it arises in practice, is proved in the course of the work and needs no special axiom. The fundamental relation is transitive; we express this as follows.

AXIOM 2. *If  $p$ ,  $q$ ,  $r$ ,  $s$  are four propositions, and, given  $p$ ,  $q$  is more probable than  $r$  and  $r$  is more probable than  $s$ , then, given  $p$ ,  $q$  is more probable than  $s$ .*

The extreme degrees of probability are certainty and impossibility. These lead to

AXIOM 3. *All propositions deducible from a proposition  $p$  have the same probability on data  $p$ ; and all propositions inconsistent with  $p$  have the same probability on data  $p$ .*

We need this axiom to ensure consistency with deductive logic in cases that can be treated by both methods. We are trying to construct an extended logic, of which deductive logic will be a part, not to introduce an ambiguity in cases where deductive logic already gives definite answers. I shall often speak of 'certainty on data  $p$ ' and 'impossibility on data  $p$ '. These do not refer to the mental certainty of any particular individual, but to the relations of deductive logic expressed by ' $q$  is deducible from  $p$ ' and ' $\text{not-}q$  is deducible from  $p$ '. In G. E. Moore's terminology, we may read the former as ' $p$  entails  $q$ '. In consequence of our rule 5, we shall never have ' $p$  entails  $q$ ' if  $p$  is merely the general rules of the theory and  $q$  is an empirical proposition.

Actually I shall take 'entails' in a slightly extended sense; in some usages it would be held that  $p$  is not deducible from  $p$ , or from  $p$  and  $q$  together. Some shortening of the writing is achieved if we agree to define ' $p$  entails  $q$ ' as meaning either ' $q$  is deducible from  $p$ ' or ' $q$  is identical with  $p$ ' or ' $q$  is identical with some proposition asserted in  $p$ '. This avoids the need for special attention to trivial cases.

We also need the following axiom.

AXIOM 4. *If, given  $p$ ,  $q$  and  $q'$  cannot both be true, and if, given  $p$ ,*

*r* and *r'* cannot both be true, and if, given *p*, *q* and *r* are equally probable and *q'* and *r'* are equally probable, then, given *p*, '*q* or *q'*' and '*r* or *r'*' are equally probable.

At this stage it is desirable for clearness to introduce the following notations and terminologies, mainly from *Principia Mathematica*.

$\sim p$  means 'not-*p*'; that is, *p* is false.

$p.q$  means '*p* and *q*'; that is, *p* and *q* are both true.

$p \vee q$  means '*p* or *q*'; that is, at least one of *p* and *q* is true.

These notations may be combined, dots being used as brackets. Thus

$\sim : p.q$  means '*p* and *q* is not true'; that is, at least one of *p* and *q* is false, which is equivalent to  $\sim p \vee \sim q$ . But

$\sim p.q$  means '*p* is false and *q* is true', which is not the same proposition. The rule is that a set of dots represents a bracket, the completion of the bracket being either the next equal set of dots or the end of the expression. Dots may be omitted in joint assertions where no ambiguity can arise.

The *joint assertion* or *conjunction* of *p* and *q* is the proposition  $p.q$ ; and the joint assertion of *p*, *q*, *r*, *s*,... is the proposition  $p.q.r.s...$ ; that is, that *p*, *q*, *r*, *s*,... are all true. The joint assertion is also called the *logical product*.

The *disjunction* of *p* and *q* is the proposition  $p \vee q$ ; the disjunction of *p*, *q*, *r*, *s* is the proposition  $p \vee q \vee r \vee s$ , that is, at least one of *p*, *q*, *r*, *s* is true. The disjunction is also called the *logical sum*.

A set of propositions  $q_i$  ( $i = 1$  to  $n$ ) are said to be *exclusive* on data *p* if not more than one of them can be true on data *p*; that is, if *p* entails all the disjunctions  $\sim q_i \vee \sim q_k$  when  $i \neq k$ .

A set of propositions *q*, *r*, *s* are said to be *exhaustive* on data *p* if at least one of them must be true on data *p*; that is, if *p* entails the disjunction  $q \vee r \vee s$ .

It is possible for a set of alternatives to be both exclusive and exhaustive. For instance, a finite class must have some number *n*; then the propositions  $n = 0, 1, 2, 3, \dots$  must include one true proposition, but cannot contain more than one.

Then Axiom 4 will read:

*If q and q' are exclusive, and r and r' are exclusive, on data p, and if, given p, q and r are equally probable and q' and r' are equally probable, then, given p, q ∨ q' and r ∨ r' are equally probable.*

An immediate extension, obtained by successive applications of this axiom, is:

**THEOREM 1.** *If  $q_1, q_2, \dots, q_n$  are exclusive, and  $r_1, r_2, \dots, r_n$  are exclusive, on data  $p$ , and if, given  $p$ , the propositions  $q_1$  and  $r_1$ ,  $q_2$  and  $r_2$ , ...,  $q_n$  and  $r_n$  are equally probable in pairs, then given  $p$ ,  $q_1 \vee q_2 \dots \vee q_n$  and  $r_1 \vee r_2 \dots \vee r_n$  are equally probable.*

It will be noticed that we have not yet assumed that probabilities can be expressed by numbers. I do not think that the introduction of numbers is strictly necessary to the further development; but it has the enormous advantage that it permits us to use mathematical technique. Without it, while we might obtain a set of propositions that would have the same meanings, their expression would be much more cumbrous. The actual introduction of numbers is done by conventions, the nature of which is essentially linguistic.

**CONVENTION 1.** *We assign the larger number on given data to the more probable proposition (and therefore equal numbers to equally probable propositions).*

**CONVENTION 2.** *If, given  $p$ ,  $q$  and  $q'$  are exclusive, then the number assigned on data  $p$  to ' $q$  or  $q'$ ' is the sum of those assigned to  $q$  and to  $q'$ .*

It is important to notice the meaning of a convention. It is neither an axiom nor a theorem. It is merely a rule introduced for convenience, and it has the property that other rules would give the same results. W. E. Johnson remarks that a convention is properly expressed in the imperative mood. An instance is the use of rectangular or polar coordinates in Euclidean geometry. The distance between two points is the fundamental idea, and all propositions can be stated as relations between distances. Any proposition in rectangular coordinates can be translated into polar coordinates, or vice versa, and both expressions would give the same results if translated into propositions about distances. It is purely a matter of convenience which we choose in a particular case. The choice of a unit is always a convention. But care is needed in introducing conventions; some postulate of consistency about the fundamental ideas is liable to be hidden. It is quite easy to define an equilateral right-angled plane triangle, but that does not make such a triangle possible. In this case Convention 1 specifies what order the numbers are to be arranged in. Numbers can be arranged in an order, and so can probabilities, by Axioms 1 and 2. The relation 'greater than' between numbers is transitive, and so is the relation 'more probable than' between propositions on the same data. Therefore it is possible to assign numbers by Convention 1, so that the order of increasing degrees of belief will be the order of increasing number.



So far we need no new axiom; but we shall need the axiom that there are enough numbers for our purpose.

**AXIOM 5.** *The set of possible probabilities on given data, ordered in terms of the relation 'more probable than', can be put into one-one correspondence with a set of real numbers in increasing order.*

The need for such an axiom was pointed out by an American reviewer of *Scientific Inference*. He remarked that if we take a series of number pairs  $u_n = (a_n, b_n)$  and make it a rule that  $u_r$  is to be placed after  $u_s$  if  $a_r > a_s$ , but that if  $a_r = a_s$ ,  $u_r$  is to be placed after  $u_s$  if  $b_r > b_s$ , then the axiom that the  $u_n$  can be placed in an order will hold, but if  $a_n$  and  $b_n$  can each take a continuous series of values it will be impossible to establish a one-one correspondence between the pairs and a single continuous series without deranging the order.

Convention 2 and Axiom 4 will imply that, if we have two pairs of exclusive propositions with the same probabilities on the same data, the numbers chosen to correspond to their disjunctions will be the same. The extension to disjunctions of several propositions is justified by Theorem 1. We shall always, on given data, associate the same numbers with propositions entailed or contradicted by the data; this is justified by Axiom 3. The assessment of numbers in the way suggested is therefore consistent with our axioms. We can now introduce the formal notation

$$P(q | p)$$

for the number associated with the probability of the proposition  $q$  on data  $p$ ; it may be read 'the probability of  $q$  given  $p$ ' provided that we remember that the number is not in fact the probability, but merely a representation of it in terms of a pair of conventions. The probability, strictly, is the reasonable degree of confidence and is not identical with the number used to express it. The relation is that between Mr. Smith and his name 'Mr. Smith'. A sentence containing the words 'Mr. Smith' may correspond to, and identify, a fact about Mr. Smith. But Mr. Smith himself does not occur in the sentence.† In this notation, the properties of numbers will now replace Axiom 1; Axiom 2 is restated

'if  $P(q | p) > P(r | p)$ , and  $P(r | p) > P(s | p)$ , then  $P(q | p) > P(s | p)$ ', which is a mere mathematical implication, since all the expressions are numbers. Axiom 3 will require us to decide what numbers to associate with certainty and impossibility. We have

**THEOREM 2.** *If  $p$  is consistent with the general rules, and  $p$  entails  $\sim q$ , then  $P(q | p) = 0$ .*

† Cf. R. Carnap, *The Logical Syntax of Language*.

For let  $q$  and  $r$  be any two propositions, both impossible on data  $p$ . Then (Ax. 3) if  $a$  is the number associated with impossibility on data  $p$ ,

$$P(q | p) = P(r | p) = P(q \vee r | p) = a$$

since  $q$ ,  $r$ , and  $q \vee r$  are all impossible propositions on data  $p$  and must be associated with the same number. But  $qr$  is impossible on data  $p$ ; hence, by definition,  $q$  and  $r$  are exclusive on data  $p$ , and (Conv. 2)

$$P(q \vee r | p) = P(q | p) + P(r | p) = 2a;$$

whence  $a = 0$ . Therefore all probability numbers are  $\geq 0$ , by Convention 1.

As we have not assumed the comparability of probabilities on different data, attention is needed to the possible forms that can be substituted for  $q$  and  $r$ , given  $p$ . If  $p$  is a purely *a priori* proposition, it can never entail an empirical one. Hence, if  $p$  stands for our general rules, the admissible values for  $q$  and  $r$  must be false *a priori* propositions, such as  $2 = 1$  and  $3 = 2$ . Since such propositions can be stated the theorem follows. If  $p$  is empirical, then  $\sim p$  is an admissible value for both  $q$  and  $r$ . Or, since we are maintaining the same general principles throughout, we may remember that in practice if  $p$  is empirical and we denote our general principles by  $h$ , then any set of data that actually occurs and includes an empirical proposition will be of the form  $ph$ . Then for  $q$  and  $r$  we may still substitute false *a priori* propositions, which will be impossible on data  $ph$ . Hence it is always possible to assign  $q$  and  $r$  so as to satisfy the conditions stated in the proof.

CONVENTION 3. If  $p$  entails  $q$ , then  $P(q | p) = 1$ .

This is the rule generally adopted; but there are cases where we wish to express ignorance over an infinite range of values of a quantity, and it is then convenient to express certainty that the quantity lies in that range by  $\infty$ , in order to keep ratios for finite ranges determinate. None of our axioms so far has stated that we must always express certainty by the same number on different data, merely that we must on the same data; but with this exception it is convenient to do so.

The converse of Theorem 2 would be: 'If  $P(q | p) = 0$ , then  $p$  entails  $\sim q$ .' This is false if we use Convention 3. For instance, a continuous variable may be equally likely to have any value between 0 and 1. Then the probability that it is exactly  $\frac{1}{2}$  is 0, but  $\frac{1}{2}$  is not an impossible value. There would be no point in making certainty correspond to infinity in such a case, for it would make the probability infinite for

any finite range. It turns out that we have no occasion to use the converse of Theorem 2.

AXIOM 6. *If  $pq$  entails  $r$ , then  $P(qr | p) = P(q | p)$ .*

In other words, given  $p$  throughout, we may consider whether  $q$  is false or true. If  $q$  is false, then  $qr$  is false. If  $q$  is true, then, since  $pq$  entails  $r$ ,  $r$  is also true and therefore  $qr$  is true. Similarly, if  $qr$  is true it entails  $q$ , and if  $qr$  is false  $q$  must be false on data  $p$ , since if it was true  $qr$  would be true. Thus it is impossible, given  $p$ , that either  $q$  or  $qr$  should be true without the other. This is an extension of Axiom 3 and is necessary to enable us to take over a further set of rules suggested by deductive logic, and to say that all equivalent propositions have the same probability on given data.

THEOREM 3. *If  $q$  and  $r$  are equivalent in the sense that each entails the other, then each entails  $qr$ , and the probabilities of  $q$  and  $r$  on any data must be equal. Similarly, if  $pq$  entails  $r$ , and  $pr$  entails  $q$ ,  $P(q | p) = P(r | p)$ , since both are equal to  $P(qr | p)$ .*

An immediate corollary is

THEOREM 4.  $P(q | p) = P(qr | p) + P(q \cdot \sim r | p)$ .

For  $qr$  and  $q \cdot \sim r$  are exclusive, and the sum of their probabilities on any data is the probability of  $qr \vee q \cdot \sim r$  (Conv. 2). But  $q$  entails this proposition, and also, if either  $q$  and  $r$  are both true or  $q$  is true and  $r$  false,  $q$  is true in any case. Hence the propositions  $q$  and  $qr \vee q \cdot \sim r$  are equivalent, and the theorem follows by Theorem 3.

It follows further that  $P(q | p) \geq P(qr | p)$ , since  $P(q \cdot \sim r | p)$  cannot be negative. Also, if we write  $q \vee r$  for  $q$ , we have

$$P(q \vee r | p) = P(q \vee r : r | p) + P(q \vee r : \sim r | p) \quad (\text{Th. 4})$$

and  $q \vee r : r$  is equivalent to  $r$ , and  $q \vee r : \sim r$  to  $q \cdot \sim r$ . Hence

$$P(q \vee r | p) \geq P(r | p).$$

THEOREM 5. *If  $q$  and  $r$  are two propositions, not necessarily exclusive on data  $p$ ,*

$$P(q | p) + P(r | p) = P(q \vee r | p) + P(qr | p).$$

For the propositions  $qr$ ,  $q \cdot \sim r$ ,  $\sim q \cdot r$ ,  $\sim q \cdot \sim r$  are exclusive; and  $q$  is equivalent to the disjunction of  $qr$  and  $q \cdot \sim r$ , and  $r$  to the disjunction of  $qr$  and  $\sim q \cdot r$ . Hence the left side of the equation is equal to

$$2P(qr | p) + P(q \cdot \sim r | p) + P(\sim q \cdot r | p) \quad (\text{Th. 4}).$$

Also  $q \vee r$  is equivalent to the disjunction of  $qr$ ,  $q \cdot \sim r$ , and  $\sim q \cdot r$ .

Hence

$$P(q \vee r | p) = P(qr | p) + P(q \cdot \sim r | p) + P(\sim q \cdot r | p) \quad (\text{Th. 4}),$$

whence the theorem follows.

It follows that, whether  $q$  and  $r$  are exclusive or not,

$$P(q \vee r | p) \leq P(q | p) + P(r | p),$$

since  $P(qr | p)$  cannot be negative. Theorems 4 and 5 together express upper and lower bounds to the possible values of  $P(q \vee r | p)$  irrespective of exclusiveness. It cannot be less than either  $P(q | p)$  or  $P(r | p)$ ; it cannot be more than  $P(q | p) + P(r | p)$ .

**THEOREM 6.** *If  $q_1, q_2, \dots$  are a set of equally probable and exclusive alternatives on data  $p$ , and if  $Q$  and  $R$  are disjunctions of two subsets of these alternatives, of numbers  $m$  and  $n$ , then  $P(Q | p)/P(R | p) = m/n$ .*

For if  $a$  is any one of the equal numbers  $P(q_1 | p), P(q_2 | p), \dots$  we have, by Convention 2,

$$P(Q | p) = ma; \quad P(R | p) = na;$$

whence the theorem follows.

**THEOREM 7.** *In the conditions of Theorem 6, if  $q_1, q_2, \dots, q_n$  are exhaustive on data  $p$ , and  $R$  denotes their disjunction, then  $R$  is entailed by  $p$  and*

$$P(R | p) = 1 \quad (\text{Conv. 3}).$$

*It follows that  $P(Q | p) = m/n$ .*

This is virtually Laplace's rule, stated at the opening of the *Théorie Analytique*.  $R$  entails itself and therefore is a possible value of  $p$ ; hence

$$P(Q | R) = m/n.$$

This may be read: *given that a set of alternatives are equally probable, exclusive, and exhaustive, the probability that some one of any subset is true is the ratio of the number in that subset to the whole number of possible cases.* This form depends on Convention 3, and must be used only in cases where that convention is adopted. Theorem 6, however, is independent of Convention 3. If we chose to express certainty on data  $p$  by 2 instead of 1, the only change would be that all numbers associated with probabilities on data  $p$  would be multiplied by 2, and Theorem 6 would still hold. Theorem 6 is also consistent with the possibility that the number of alternatives is infinite, since it requires only that  $Q$  and  $R$  shall be finite subsets. But in this case the number associated with the probability of any infinite subset may be infinite and Convention 3 is then unsuitable.

Theorems 6 and 7 tell us how to assess the ratios of probabilities, and, subject to Convention 3, the actual values, provided that the propositions considered can be expressed as finite subsets of equally probable, exclusive, and, for Theorem 7, exhaustive alternatives on the data. Such assessments will always be rational fractions, and may be called *R*-probabilities. Now a statement that *m* and *n* cannot exceed some given value would be an empirical proposition asserted *a priori*, and would be inadmissible on rule 5. Hence the *R*-probabilities possible within the formal scheme form a set of the ordinal type of the rational fractions.

If all probabilities were *R*-probabilities there would be no need for Axiom 5, and the converse of Theorem 2 could hold. But many propositions that we shall have to consider are of the form that a magnitude, capable of a continuous range of values, lies within a specified part of that range, and we may be unable to express them in the required form. Thus there is no need for all probabilities to be *R*-probabilities. However, if a proposition is not expressible in the required form, it will still be associated with a reasonable degree of belief by Axiom 1, and this, by Axiom 2, will separate the degrees for *R*-probabilities into two segments, according to the relations 'more probable than' and 'less probable than'. The corresponding numbers, the *R*-probabilities themselves, will be separated by a unique real number, by Axiom 5 and an application of Dedekind's section. We take the numerical assessment of the probability of a proposition not expressible in the form required by Theorems 6 and 7 to be this number. Hence we have

**THEOREM 8.** *Any probability can be expressed by a real number.*

If *x* is a variable capable of a continuous set of values, we may consider the probability on data *p* that *x* is less than *x*<sub>0</sub>, say

$$P(x < x_0 | p) = f(x_0).$$

If *f*(*x*<sub>0</sub>) is differentiable we shall then be able to write

$$P(x_0 < x < x_0 + dx_0 | p) = f'(x_0)dx_0 + o(dx_0).$$

We shall usually write this briefly  $P(dx | p) = f'(x)dx$ , *dx* on the left meaning the proposition that *x* lies in a particular range *dx*. *f'*(*x*) is called the *probability density*.

**THEOREM 9.** *If Q is the disjunction of a set of exclusive alternatives on data p, and if R and S are subsets of Q (possibly overlapping) and if*

the alternatives in  $Q$  are all equally probable on data  $p$  and also on data  $Rp$ , then

$$P(RS|p) = P(R|p)P(S|Rp)/P(R|Rp).$$

For suppose that the propositions contained in  $Q$  are of number  $n$ , that the subset  $R$  contains  $m$  of them, and that the part common to  $R$  and  $S$  contains  $l$  of them. Put

$$P(Q|p) = a.$$

Then, by Theorem 6,

$$P(R|p) = ma/n; \quad P(RS|p) = la/n.$$

$P(S|Rp)$  is the probability that the true proposition is in the  $S$  subset given that it is in the  $R$  subset and  $p$ , and therefore is equal to  $(l/m)P(R|Rp)$ . Also  $RS$  entails  $R$ ; hence

$$P(S|Rp) = P(SR|Rp) \quad (\text{Ax. 6})$$

and

$$P(RS|p) = (l/m)(ma/n) = P(R|p)P(S|Rp)/P(R|Rp).$$

This is the first proposition that we have had that involves probabilities on different data, two of the factors being on data  $p$  and two on data  $Rp$ .  $Q$  itself does not appear in it and is therefore irrelevant. It is introduced into the theorem merely to avoid the use of Convention 3. It might be identical with any finite set that includes both  $R$  and  $S$ .

The proof has assumed that the alternatives considered are equally probable both on data  $p$  and also on data  $Rp$ . It has not been found possible to prove the theorem without using this condition. But it is necessary to further developments of the theory that we shall have some way of relating probabilities on different data, and Theorem 9 suggests the simplest general rule that they can follow if there is one at all. We therefore take the more general form as an axiom, as follows.

AXIOM 7. For any propositions  $p, q, r$ ,

$$P(qr|p) = P(q|p)P(r|qp)/P(q|qp).$$

If we use Convention 3 on data  $qp$  (not necessarily on data  $p$ ),  $P(q|qp) = 1$ , and we have W. E. Johnson's form of the *product rule*, which can be read: *the probability of the joint assertion of two propositions on any data  $p$  is the product of the probability of one of them on data  $p$  and that of the other on the first and  $p$ .*

We notice that the probability of the logical sum follows the addition rule (with a caveat), that of the logical product the product rule. This parallel between the *Principia* and probability language is lost when the joint assertion is called the sum, as has occurred in some recent writings.

In a sense a probability can be regarded as a logical quotient, since in the conditions of Theorem 7 the probability of  $Q$  given  $R$  is the probability of  $Q$  given  $p$  divided by that of  $R$  given  $p$ . This has been recognized in the history of the notation, which Keynes† traces to H. McColl. McColl wrote the probability of  $a$ , relative to the *a priori* premiss  $h$ , as  $a/\epsilon$ , and relative to  $bh$  as  $a/b$ . This was modified by W. E. Johnson to  $a/h$  and  $a/bh$ , and he is followed by Keynes, Broad, and Ramsey. Wrinch and I found that this notation was inconvenient when the solidus may have to be used in its usual mathematical sense in the same equation, and introduced  $P(p:q)$ , which I modified further to  $P(p|q)$  in *Scientific Inference* because the colon was beginning to be needed in the *Principia* sense of a bracket.

The sum of two classes  $\alpha$  and  $\beta$ , in *Principia*, is the class  $\gamma$  such that every member of  $\alpha$  or of  $\beta$  is in  $\gamma$ , and conversely. The product class of  $\alpha$  and  $\beta$  is the class  $\delta$  of members common to  $\alpha$  and  $\beta$ . Thus Theorem 5 has a simple analogy with the numbers of members of the classes  $\alpha$  and  $\beta$ ,  $\gamma$  and  $\delta$ . The multiplicative class of  $\alpha$  and  $\beta$  is the class of all pairs, one from  $\alpha$  and one from  $\beta$ ; it is this class, not the product class, that gives an interpretation to the product of the *numbers* of members of  $\alpha$  and  $\beta$ .

The extension of the product rule from Theorem 9 to Axiom 7 has been taken as axiomatic. This is an application of a principle repeatedly adopted in *Principia Mathematica*. If there is a choice between possible axioms, we take the one that enables most consequences to be drawn. Such a generalization is not inductive. What we are doing is to seek for a set of axioms that will permit the construction of a theory of induction, the axioms themselves being primitive postulates. The choice is limited by rule 6; the axioms must be reduced to the minimum number, and the check on whether we make them too general will be provided by rule 2, which will reject a theory if it is found to lead to contradictory consequences. Consider then whether the rule

$$P(qr|p) = P(q|p)P(r|qp)$$

can hold in general. Suppose first that  $p$  entails  $\sim:qr$ ; then either  $p$  entails  $\sim q$ , or  $p$  and  $q$  together entail  $\sim r$ . In either case both sides of the equation vanish and the rule holds. Secondly, suppose that  $p$  entails  $qr$ ; then  $p$  entails  $q$  and  $pq$  entails  $r$ . Thus both sides of the equation are 1. Similarly, we have consistency in the converse cases where  $p$

† *Treatise on Probability*, 1921, p. 155. This book is full of interesting historical data and contains many important critical remarks. It is not very successful on the constructive side, since an unwillingness to generalize the axioms has prevented Keynes from obtaining many important results.

entails  $\sim q$ , or  $pq$  entails  $\sim r$ , or  $p$  entails  $q$  and  $pq$  entails  $r$ . This covers the extreme cases.

If there are any cases where the rule is untrue, we shall have to say that in such cases  $P(qr | p)$  depends on something besides  $P(q | p)$  and  $P(r | qp)$ , and a new hypothesis would be needed to deal with such cases. By rule 6, we must not introduce any such hypothesis unless need for it is definitely shown. The product rule may therefore be taken as general unless it can be shown to lead to contradictions. We shall see (p. 35) that consistency can be proved in a wide class of cases.

**1.21.** The product rule is often misread as follows: the joint probability of two propositions is the product of their probabilities separately. This is meaningless as it stands because the data relative to which the probabilities are considered are not mentioned. In actual application, the rule so stated is liable to become: the joint probability of two propositions on given data is the product of their separate probabilities on those data. This is false. We may see this by considering extreme cases. The correct statement of the rule may be written (using Convention 3 on data  $pr$ )

$$P(pq | r) = P(p | r)P(q | pr) \quad (1)$$

and the other one as

$$P(pq | r) = P(p | r)P(q | r). \quad (2)$$

If  $p$  cannot be true given  $r$ , then  $p$  and  $q$  cannot both be true, and both (1) and (2) reduce to  $0 = 0$ . If  $p$  is certain given  $r$ , both reduce to

$$P(q | r) = P(q | r) \quad (3)$$

since in (1) the inclusion of  $p$  in the data tells us nothing about  $q$  that is not already told us by  $r$ . If  $q$  is impossible given  $r$ , both reduce to  $0 = 0$ . If  $q$  is certain given  $r$ , both reduce to

$$P(p | r) = P(p | r). \quad (4)$$

So far everything is satisfactory. But suppose that  $q$  is impossible given  $pr$ . Then it is impossible for  $pq$  to be true given  $r$ , and (1) reduces correctly to  $0 = 0$ . But (2) reduces to

$$0 = P(p | r)P(q | r),$$

which is false; it is perfectly possible for both  $p$  and  $q$  to be consistent with  $r$  and  $pq$  to be inconsistent with  $r$ . Consider the following. Let  $r$  consist of the following information: in a given population all the members have eyes of the same colour; half of them have blue eyes and half brown; one member is to be chosen, and any member is equally likely to be selected.  $p$  is the proposition that his left eye is blue,  $q$  the



proposition that his right eye is brown. What is the probability, on data  $r$ , that his left eye is blue and his right brown?  $P(p|r)$  and  $P(q|r)$  are both  $\frac{1}{2}$ , and according to (2)  $P(pq|r) = \frac{1}{4}$ . But according to (1) the probability that his right eye is brown must be assessed subject both to the information that his eyes are of the same colour and that his left eye is blue, and this probability is 0. Thus (1) gives  $P(pq|r) = 0$ . Clearly the latter result is right; further applications of the former, considering also  $\sim p$  (left eye brown) and  $\sim q$  (right eye blue) lead to the astonishing result that on data including the proposition that all members have two eyes of the same colour, it is as likely as not that any member will have eyes of different colours.

This trivial instance is enough to dispose of (2); but (2) has been widely applied in cases where it gives wrong results, and sometimes seriously wrong ones. The Boltzmann  $H$ -theorem of the kinetic theory of gases rests on a fallacious application of it, since it considers an assembly of molecules, possibly with differences of density from place to place, and gives the joint probability that two molecules will be in adjoining regions as the product of the separate probabilities that they will be. If there are differences of density, and one molecule is in a region chosen at random, that is some evidence that the region is one of high density; then the probability that a second is in the region, given that the first is, is somewhat higher than it would be in the absence of information about the first. Similar considerations apply to Boltzmann's treatment of the velocities. In this case the mistake has not prevented the right result from being obtained, though it does not follow from the hypotheses.

Nevertheless there are many cases where (2) is true. If

$$P(q|pr) = P(q|r)$$

we say that  $p$  is *irrelevant* to  $q$ , given  $r$ .

**1.22. THEOREM 10.** *If  $q_1, q_2, \dots, q_n$  are a set of alternatives,  $H$  the information already available, and  $p$  some additional information, then the ratio*

$$\frac{P(q_r|pH)P(q_r|q_rH)}{P(q_r|H)P(p|q_rH)}$$

*is the same for all the  $q_r$ .*

By Axiom 7

$$P(pq_r|H) = P(p|H)P(q_r|pH)/P(p|pH) \quad (1)$$

$$= P(q_r|H)P(p|q_rH)/P(q_r|q_rH), \quad (2)$$

whence 
$$\frac{P(q_r | pH)P(q_r | q_r H)}{P(q_r | H)P(p | q_r H)} = \frac{P(p | pH)}{P(p | H)} \quad (3)$$

which is independent of  $q_r$ .

If we use unity to denote certainty on data  $q_r H$  for all the  $q_r$ , (3) becomes 
$$P(q_r | pH) \propto P(q_r | H)P(p | q_r H) \quad (4)$$

for variations of  $q_r$ . This is the *principle of inverse probability*, first given by Bayes in 1763. It is the chief rule involved in the process of learning from experience. It may also be stated, by means of the product rule, as follows:

$$P(q_r | pH) \propto P(pq_r | H). \quad (5)$$

This is the form used by Laplace, by way of the statement that the posterior probabilities of causes are proportional to the probabilities *a priori* of obtaining the data by way of those causes. In the form (4), if  $p$  is a description of a set of observations and the  $q_r$  a set of hypotheses, the factor  $P(q_r | H)$  may be called the *prior probability*,  $P(q_r | pH)$  the *posterior probability*, and  $P(p | q_r H)$  the *likelihood*, a convenient term introduced by Professor R. A. Fisher, though in his usage it is sometimes multiplied by a constant factor. It is the probability of the observations given the original information and the hypothesis under discussion. The term *a priori* probability is sometimes used for the prior probability, but this term has been used in so many senses that the only solution is to abandon it. To Laplace the *a priori* probability meant  $P(pq_r | H)$ , and sometimes the term has even been used for the likelihood. *A priori* has a definite meaning in logic, in relation to propositions independent of experience, and we frequently have need to use it in this sense. We may then state the principle of inverse probability in the form: *The posterior probabilities of the hypotheses are proportional to the products of the prior probabilities and the likelihoods.* The constant factor will usually be fixed by the condition that one of the propositions  $q_1$  to  $q_n$  must be true, and the posterior probabilities must therefore add up to 1. (If 1 is not suitable to denote certainty on data  $pH$ , no finite set of alternatives will contain a finite fraction of the probability. The rule covers all cases when there is anything to say.)

The use of the principle is easily seen in general terms. If there is originally no ground to believe one of a set of alternatives rather than another, the prior probabilities are equal. The most probable, when evidence is available, will then be the one that was most likely to lead to that evidence. We shall be most ready to accept the hypothesis that

requires the fact that the observations have occurred to be the least remarkable coincidence. On the other hand, if the data were equally likely to occur on any of the hypotheses, they tell us nothing new with respect to their credibility, and we shall retain our previous opinion, whatever it was. The principle will deal with more complicated circumstances also; the immediate point is that it does provide us with what we want, a formal rule in general accordance with common sense, that will guide us in our use of experience to decide between hypotheses.

1.23. We have not yet shown that Convention 2 is a convention and not a postulate. This must be done by considering other possible conventions and seeing what results they lead to. Any other convention must not contradict Axiom 4. For instance, if the number associated with a probability by our rules is  $x$ , we might agree instead to use the number  $e^x$ . Then if  $x$  and  $x'$  are the present estimates for the propositions  $q$  and  $q'$ , and for  $r$  and  $r'$ , those for  $q \vee q'$  and  $r \vee r'$  will both be  $e^{x+x'}$  and the consistency rule of Axiom 4 will be satisfied. But instead of the addition rule for the number to be associated with a disjunction we shall have a product rule. Every proposition stated in either notation can be translated into the other; if our present system leads to the result that a hypothesis is as likely to be true as it is that we should pick a white ball at random out of a bag containing 99 white ones and 1 black one, that result will also be obtained on the suggested alternative system. The fundamental notion is that of the comparison of reasonable degrees of belief, and so long as all methods place them in the same order the differences between the methods are conventional. This will be satisfied if instead of the number  $x$  we choose any function of it,  $f(x)$ , such that  $x$  and  $f(x)$  are increasing functions of each other, so that for any value of one the other is determinate. This is necessary by Convention 1 and Axiom 1, but every form of  $f(x)$  will lead to a different rule for the probability-number of a disjunction if it is to be consistent with Axiom 4. Hence the addition rule is a convention. It is, of course, much the easiest convention to use. To abandon Convention 1, consistently with Axiom 1, would merely arrange all numerical assessments in the opposite order, and again the same results would be obtained in translation. The assessment by numbers is simply a choice of the most convenient language for our purposes.

1.3. The original development of the theory, by Bayes,<sup>†</sup> proceeds differently. The foregoing account is entirely in terms of rules for the

<sup>†</sup> *Phil. Trans.* 53, 1763, 376–98.

comparison of reasonable degrees of belief. Bayes, however, takes as his fundamental idea that of expectation of benefit. This is partly a matter of what we want, which is a separate problem from that of what it is reasonable to believe; I have therefore thought it best to proceed as far as possible in terms of the latter alone. Nevertheless, we have in practice often to make decisions that involve not only belief but the desirability of the possible effect of different courses of action. If we have to give advice to a practical man, either we or he must take these into account. In deciding on his course of action he must allow both for the probability that the action chosen will lead to a certain result and for the value to him of that result if it happens. The fullest development on these lines is that of F. P. Ramsey.† I shall not attempt to reproduce it, but shall try to indicate some of the principal points as they occur in his work or in Bayes's. The fundamental idea is that the values of expectations of benefit can be arranged in an order; it is legitimate to compare a small probability of a large gain with a large probability of a small gain. The idea is necessarily more complicated than my Axiom 1; on the other hand, the comparison is one that a business man often has to make, whether he wants to or not, or whether it is legitimate or not. The rule simply says that in given circumstances there is always a best way to act. The comparison of probabilities follows at once; if the benefits are the same, whichever of two events happens, then if the values to us of the expectations of benefit differ it is because the events are not equally likely to happen, and the larger value is associated with the larger probability. Now we have to consider the combination of expectations. Here Bayes, I think, overlooks the distinction between what Laplace calls 'mathematical' and 'moral' expectation. Bayes speaks in terms of monetary stakes, and would say that a  $1/100$  chance of receiving £100 is as valuable as a certainty of receiving £1. A gambler might say that it is more valuable; most people would perhaps say that it is less so. Indeed Bayes's *definition* of a probability of  $1/100$  would be that it is the probability such that the value of the chance of receiving £100 is the same as the value of a certain £1. Since different values may be compared, the uniqueness of a probability so defined requires a postulate that the value of the expectation, the proposition and the data remaining the same, is proportional to the value to be received if the proposition

† *The Foundations of Mathematics*, 1931, pp. 157–211. This essay, like that of Bayes, was published after the author's death, and suffers from a number of imperfections in the verbal statement that he might have corrected.

is true. This is taken for granted by Bayes, and Ramsey makes an equivalent statement (foot of p. 179). The difficulty is that the value of £1 to us depends on how much money we have already. This point was brought out by Daniel Bernoulli in relation to what was called the Petersburg Problem. Two players play according to the following rules. A coin is to be thrown until a head is thrown. If it gives a head on the first throw, *A* is to pay *B* £1; if the first head is on the second throw, £2; on the third, £4; and so on. What is the fair sum for *B* to pay *A* for his chances? The mathematical expectation in pounds is

$$\frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{8} \cdot 4 + \frac{1}{16} \cdot 8 + \dots = \infty.$$

Thus on this analysis *B* should pay *A* an infinite sum. If we merely consider a large finite sum, such as £2<sup>20</sup>, he will lose if there is a head in any of the first 20 throws; he will gain considerably if the first head is on the 21st or a later throw. The question was, is it really worth anybody's while to risk such a sum, most of which he is practically certain to lose, for an almost inappreciable chance of an enormous gain? Even eighteenth-century gamblers seem to have had doubts about it. Daniel Bernoulli's solution was that the value of £2<sup>20</sup> is very different according to the amount we have to start with. The value of a loss of that sum to anybody that has just that amount is not equal and opposite to the value of a gain of the same sum. He suggested a law relating the value of a gain to the amount already possessed, which need not detain us;† but the important point is that he recognized that expectations of benefit are not necessarily additive. What Laplace calls 'moral expectation' is the value or pleasure to us of an event; its relation to the monetary value in terms of mathematical expectation may be rather remote. Bayes wrote after Bernoulli, but before Laplace, but he does not mention Bernoulli. Nevertheless, the distinction does not dispose of the interest of the treatment in terms of expectation of benefit. Though we cannot regard the benefits of gains of the same kind as mutually irrelevant, on account of this psychological phenomenon of satiety, there do seem to be many cases where benefits are mutually irrelevant. For instance, the pleasures to me of two dinners on consecutive nights seem to be nearly independent, though those of two dinners on the same night are definitely not. The pleasures of the unexpected return of a loan, having a paper accepted for publication, a swim in the afternoon, and a theatre in the evening do seem

† It is that the value of a gain  $dx$ , when we have  $x$  already, is proportional to  $dx/x$ ; this is the rule associated in certain biological applications with the names of Weber and Fechner.

independent. If there are a sufficient number of such benefits (or if there could be in some possible world, since all we need is consistency), a scale of the values of benefits can be constructed, which will satisfy the commutative rule of addition, and then, by Bayes's principles, one of probability in terms of them. The addition rule will then be a theorem. The product rule is treated by Bayes in the following way. We can write  $E(a, p | q)$  for the value of the expectation of receiving  $a$  if  $p$  is true, given  $q$ , and by definition of  $P(p | q)$ ,

$$E(a, p | q) = aP(p | q).$$

The proportionality of  $E(a, p | q)$  to  $a$ , given  $p$  and  $q$ , is a postulate, as we have already stated. Consider the value of the expectation of getting  $a$  if  $p$  and  $q$  are both true, given  $r$ . This is  $aP(pq | r)$ . But we may test  $p$  first and then  $q$ . If  $p$  turns out to be true, our expectation will be  $aP(q | pr)$ , since  $p$  is now among our data; if untrue, we know that we shall receive nothing. Now return to the first stage. If  $p$  is true we shall receive an expectation, whose value is  $aP(q | pr)$ , otherwise nothing. Hence our initial expectation is  $aP(q | pr)P(p | r)$ ; whence

$$P(pq | r) = P(p | r)P(q | pr).$$

Ramsey's presentation is much more elaborate, but depends on the same main ideas. The proof of the principle of inverse probability is simple. The difficulty about the separation of propositions into disjunctions of equally possible and exclusive alternatives is avoided by this treatment, but is replaced by difficulties concerning additive expectations. These are hardly practical ones in either case; no practical man will refuse to decide on a course of action merely because we are not quite sure which is the best way to lay the foundations of the theory. He assumes that the course of action that he actually chooses is the best; Bayes and Ramsey merely make the less drastic assumption that there is some course of action that is the best. In my method expectation would be defined in terms of value and probability; in theirs probability is defined in terms of values and expectations. The actual propositions are of course identical.

1.4. At any stage of knowledge it is legitimate to ask about a given hypothesis that is accepted, 'How do you know?' The answer will usually rest on some observational data. If we ask further, 'What did you think of the hypothesis before you had these data?' we may be told of some less convincing data; but if we go far enough back we shall always reach a stage where the answer must be: 'I thought the matter

worth considering, but had no opinion about whether it was true.' What was the probability at this stage? We have the answer already. If there is no reason to believe one hypothesis rather than another, the probabilities are equal. In terms of our fundamental notions of the nature of inductive inference, *to say that the probabilities are equal is a precise way of saying that we have no ground for choosing between the alternatives*. All hypotheses that are sufficiently definitely stated to give any difference between the probabilities of their consequences will be compared with the data by the principle of inverse probability; but if we do not take the prior probabilities equal we are expressing confidence in one rather than another before the data are available, and this must be done only from definite reason. To take the prior probabilities different in the absence of observational reason for doing so would be an expression of sheer prejudice. The rule that we should then take them equal is not a statement of any belief about the actual composition of the world, nor is it an inference from previous experience; it is merely the formal way of expressing ignorance. It is sometimes referred to as the Principle of Insufficient Reason (Laplace) or the equal distribution of ignorance. Bayes, in his great memoir, repeatedly says that the principle is to be used only in cases where we have no ground whatever for choosing between the alternatives. It is not a new rule in the present theory because it is an immediate application of Convention 1. Much confusion has arisen about it through misunderstanding and attempts to reinterpret it in terms of frequency definitions. My contention is that the frequency definitions themselves lead to no results of the kind that we need until the notion of reasonable degree of belief is reintroduced, and that since the whole purpose of these definitions is to avoid this notion they necessarily fail in their object. When reasonable degree of belief is taken as the fundamental notion the rule is immediate. We begin by making no assumption that one alternative is more likely than another and use our data to compare them.

Suppose that one hypothesis is suggested by one person  $A$ , and another by a dozen  $B, C, \dots$ ; does that make any difference? No; but it means that we have to attend to two questions instead of one. First, is  $p$  or  $q$  true? Secondly, is the difference between the suggestions due to some psychological difference between  $A$  and the rest? The mere voting is not evidence because it is quite possible for a large number of people to make the same mistake. The second question cannot be answered until we have answered the first, and the first must be considered on its merits apart from the second.

**1.5.** We are now in a position to consider whether we have fulfilled the conditions that we required at the outset. I think (1) is satisfied, though the history of both probability and deductive logic is a warning against over-confidence that an unstated axiom has not slipped in.

2. Axiom 1 assumes consistency, but this assumption by itself does not guarantee that a given system is consistent. It makes it possible to derive theorems by equating probabilities found in different ways, and if in spite of all efforts probabilities found in different ways were different, the axiom would make it impossible to accept the situation as satisfactory. We must not expect too much in the nature of a general proof of consistency. There is a theorem due to Gödel that if any logical system that includes arithmetic contained a proof of its own consistency, it would also contain one of its own inconsistency; so apparently it would be fatal to a system if we could find a general proof of consistency within it. Proofs of the consistency of various logical schemes (including the system of *Principia Mathematica* and therefore the theory of functions of a real variable) do exist, but only by going outside the frames of the schemes themselves. The proof amounts to finding a proposition that can be stated in the system but cannot be proved or disproved by using the rules of the system. Since the system of *Principia* contains a proposition that two contradictory propositions imply any proposition, the existence of an undemonstrable proposition implies that the primitive propositions in the system are consistent. But this argument itself cannot be expressed in *Principia* language! What we want is that the probability of a proposition on the same data shall always be the same; thus, if we are considering two alternative hypotheses  $q_1$  and  $q_2$ , our previous information is  $H$ , and the new evidence consists of two batches of data  $p_1$  and  $p_2$ , the assessments on data  $p_1 p_2 H$  should be the same whether we take  $p_1$  or  $p_2$  into account first or both at once. Now, by the principle of inverse probability,

$$\frac{P(q_1 | p_1 H)}{P(q_1 | H)P(p_1 | q_1 H)} = \frac{P(q_2 | p_1 H)}{P(q_2 | H)P(p_1 | q_2 H)}.$$

Replacing  $H$  by  $p_1 H$  and  $p_1$  by  $p_2$  we shall obtain the result for the application of the additional data  $p_2$ ,  $p_1$  being now already given:

$$\frac{P(q_1 | p_1 p_2 H)}{P(q_1 | p_1 H)P(p_2 | q_1 p_1 H)} = \frac{P(q_2 | p_1 p_2 H)}{P(q_2 | p_1 H)P(p_2 | q_2 p_1 H)}.$$

Multiplying, we have

$$\frac{P(q_1 | p_1 p_2 H)}{P(q_1 | H)P(p_1 | q_1 H)P(p_2 | p_1 q_1 H)} = \frac{P(q_2 | p_1 p_2 H)}{P(q_2 | H)P(p_1 | q_2 H)P(p_2 | p_1 q_2 H)}.$$



But by Axiom 7, assuming that the product rule holds for likelihoods,

$$P(p_1 | q_1 H) P(p_2 | p_1 q_1 H) = P(p_1 p_2 | q_1 H),$$

and therefore

$$\frac{P(q_1 | p_1 p_2 H)}{P(q_1 | H) P(p_1 p_2 | q_1 H)} = \frac{P(q_2 | p_1 p_2 H)}{P(q_2 | H) P(p_1 p_2 | q_2 H)},$$

which is the result of applying the principle of inverse probability to take account of the data  $p_1$  and  $p_2$  simultaneously. By symmetry we should obtain the same result if we took account of  $p_2$  first. Extension to any number of batches of new data is obviously possible, and the results will therefore be consistent provided that we always start with the same data and finish with the same, and that we take account of the new data as we proceed. Neglect of the last condition may lead to inconsistencies, but that is the result of not applying the principle correctly. In the proof we have assumed that the product rule holds for likelihoods. This has not been proved in general, but has invariably been assumed even by those who claim to reject the principle of inverse probability. What our theorem shows is that if the product rule holds for likelihoods the principle of inverse probability cannot lead to contradiction.

The consistency of the product rule can be treated more directly as follows. Let  $q_i, r_k$  be two sets of propositions each exclusive and exhaustive on  $p$ , and denote their disjunctions by  $Q, R$ . Then

$$P(r_k | p) = P(Q r_k | p) = \sum_i P(q_i r_k | p).$$

Instead of Axiom 7 assume that

$$\frac{P(q_i | r_k p)}{P(q_j | r_k p)} = \frac{P(q_i r_k | p)}{P(q_j r_k | p)},$$

and assume that probabilities on data  $p$  satisfy the axioms. Then for probabilities on data  $r_k p$  it is obvious that Axioms 1, 2, 5 are satisfied; Axioms 3, 4, 6, 7 are easily proved, beginning with Axiom 6. Hence if we weaken Axiom 1 to a statement that probabilities are comparable given *one* sufficiently wide datum  $p$ , we can consistently convert the product rule into a definition of probabilities on data including  $p$ .

3. For any assessment of the prior probability the principle of inverse probability will give a unique posterior probability. This can be used as the prior probability in taking account of a further set of data, and the theory can therefore always take account of new information. The choice of the prior probability at the outset, that is, before taking into account any observational information at all, requires further consideration. We shall see that further principles are available as a guide.

These principles sometimes indicate a unique choice, but in many problems some latitude is permissible, so far as we know at present. In such cases, and in a different world, the matter would be one for decision by the International Research Council. Meanwhile we need only remark that the choice in practice, within the range permitted, makes very little difference to the results.

4. This is satisfied by definition.

5. We have avoided contradicting rule 5 so far, but further applications of it will appear later.

6. Our main postulates are the existence of unique reasonable degrees of belief, which can be put in a definite order; Axiom 4 for the consistency of probabilities of disjunctions; either the axiomatic extension of the product rule or the theory of expectation. It does not appear that these can be reduced in number, without making the theory incapable of covering the ground required.

7. The simple cases mentioned on pp. 29–30 show how the principle of inverse probability does correspond to ordinary processes of learning, though we shall go into much more detail as we proceed. Differences between individual assessments that do not agree with the results of the theory will be part of the subject-matter of psychology. Their existence can be admitted without reducing the importance of a unique standard of reference. It has been said that the theory of probability could be accepted only if there was experimental evidence to support it; that psychology should invent methods of measuring actual degrees of belief and compare them with the theory. I should reply that without an impersonal method of analysing observations and drawing inferences from them we should not be in a position to interpret these observations either. The same considerations would apply to arithmetic. To quote P. E. B. Jourdain:†

‘I sometimes feel inclined to apply the historical method to the multiplication table. I should make a statistical inquiry among school children, before their pristine wisdom had been biased by teachers. I should put down their answers as to what 6 times 9 amounts to, I should work out the average of their answers to six places of decimals, and should then decide that, at the present stage of human development, this average is the value of 6 times 9.’

I would add only that without the multiplication table we should not be able to say what the average is. Nobody says that wrong answers invalidate arithmetic, and accordingly we need not say that the fact that some inferences do not agree with the theory of probability

† *The Philosophy of Mr. Bertrand Russell*, 1918, p. 88.

invalidates the theory. It is sufficiently clear that the theory does represent the main features of ordinary thought. The advantage of a formal statement is that it makes it easier to see in any particular case whether the ordinary rules are being followed.

This distinction shows that theoretically a probability should always be worked out completely. We have again an illustration from pure mathematics. What is the 1,000th figure in the expansion of  $e$ ? Nobody knows; but that does not say that the probability that it is a 5 is 0.1. By following the rules of pure mathematics we could determine it definitely, and the statement is either entailed by the rules or contradicted; in probability language, on the data of pure mathematics it is either a certainty or an impossibility.† Similarly, a guess is not a probability. Probability theory is more complicated than deductive logic, and even in pure mathematics we must often be content with approximations. Mathematical tables consist entirely of approximations. Hence we must expect that our numerical estimates of probabilities in practice will usually be approximate. The theory is in fact the system of thought of an ideal man that entered the world knowing nothing, and always worked out his inferences completely, just as pure mathematics is part of the system of thought of an ideal man who always gets his arithmetic right.‡ But that is no reason why the actual man should not do his best to approximate to it.

**1.6.** We can now indicate in general terms how an inductive inference can approach certainty, though it cannot reach it. If  $q$  is a hypothesis,  $H$  the previous information, and  $p_1$  an experimental fact, we have by two applications of the product rule, using Convention 3,

$$P(q | p_1 H) = \frac{P(q | H)P(p_1 | qH)}{P(p_1 | H)}, \quad (1)$$

since both are equal to  $P(p_1 q | H)/P(p_1 | H)$ . If  $p_1$  is a consequence of  $q$ ,  $P(p_1 | qH) = 1$ ; hence in this case

$$P(q | p_1 H) = \frac{P(q | H)}{P(p_1 | H)}. \quad (2)$$

† It is unfortunate that pure mathematicians speak of, for instance, the probability distribution of prime numbers, meaning a smoothed density distribution. Systematic botanists and zoologists are far ahead of mathematicians and physicists in tidying up their language.

‡ An expert computer does not trust his arithmetic without applying checks, which would give identities if the work is correct but would be expected to fail if there is a mistake. Thus induction is used to check the correctness of what is meant to be deduction. The possibility that two mistakes have cancelled is treated as so improbable that it can be ignored.

If  $p_1, p_2, \dots$  are further consequences of  $q$ , which are found to be true, we shall have in succession

$$P(q | p_1 p_2 H) = \frac{P(q | H)}{P(p_1 | H)P(p_2 | p_1 H)}, \quad \dots,$$

$$P(q | p_1 p_2 \dots p_n H) = \frac{P(q | H)}{P(p_1 | H)P(p_2 | p_1 H) \dots P(p_n | p_1 \dots p_{n-1} H)}. \quad (3)$$

Thus each verification divides the probability of the hypothesis by the probability of the verification, given the previous information. Thus, with a sufficient number of verifications, one of three things must happen: (1) The probability of  $q$  on the information available will exceed 1. (2) it is always 0. (3)  $P(p_n | p_1 p_2 \dots p_{n-1} H)$  will tend to 1. (1) is impossible since the highest degree of probability is certainty. (2) means that  $q$  can never reach a finite probability, however often it is verified. But if we adopt (3), repeated verifications of consequences of a hypothesis will make it practically certain that further consequences of it will be verified. This accounts for the confidence that we actually have in inductive inferences.

This proposition also provides us with an answer to various logical difficulties connected with the fact that if  $p$  entails  $q$ ,  $q$  does not necessarily entail  $p$ .  $p$  may be one of many alternatives that would also entail  $q$ . In the lowest terms, if  $q$  is the disjunction of a set of alternatives  $q_1, q_2, \dots, q_m$ , then any member of this set entails  $q$ , but  $q$  does not entail any particular member. Now in science one of our troubles is that the alternatives available for consideration are not always an exhaustive set. An unconsidered one may escape attention for centuries. The last proposition shows that this is of minor importance. It says that if  $p_1, \dots, p_n$  are successive verifications of a hypothesis  $q$ ,

$$P(p_n | p_1 p_2 \dots p_{n-1} H)$$

will approach certainty; it does not involve  $q$  and therefore holds *whether  $q$  is true or not*. The unconsidered hypothesis, if it had been thought of, would either (1) have led to the consequences  $p_1, p_2, \dots$  or (2) to different consequences at some stage. In the latter case the data would have been enough to dispose of it, and the fact that it was not thought of has done no harm. In the former case the considered and the unconsidered alternatives would have the same consequences, and will presumably continue to have the same consequences. The unconsidered alternative becomes important only when it is explicitly stated and a type of observation can be found where it would lead to different predictions from the old one. The rise into importance of the

theory of general relativity is a case in point. Even though we now know that the systems of Euclid and Newton need modification, it was still legitimate to base inferences on them until we knew what particular modification was needed. The theory of probability makes it possible to respect the great men on whose shoulders we stand.

The possibility of this procedure rests, of course, on the fact that there are cases where a large number of observations have been found to agree with predictions made by a law. The interest of an estimate of the probability of a law, given certain data, is not great unless those actually are our data. Indeed, a statement of it might lead to highly uncomplimentary remarks. It is not necessary that the predictions shall be exact. In the case of uniformly accelerated motion mentioned near the beginning, if the law is stated in the form that at any instant  $t$ , the observed  $s$  will lie between  $a + ut + \frac{1}{2}gt^2 \pm \epsilon$ , where  $\epsilon$  is small compared with the whole range of variation of  $s$ , it will still be a legitimate inference after many verifications that the law will hold in future instances within this margin of uncertainty. This takes us a further step towards understanding the nature of the acceptance of a simple law in spite of the fact that in the crude form given in applied mathematics it does not exactly agree with the observations.

**1.61.** If we lump together all hypotheses that give indistinguishable consequences, their total probability will tend to 1 with sufficient verification. For if we have a set of hypotheses  $q_1, \dots, q_m$ , all asserting that a quantity  $x$  will lie in a range  $\pm \epsilon$ , we may denote their disjunction by  $q$ , which will assert the same. Suppose that  $\sim q$  would permit the quantity to lie in a range  $\pm E$ , where  $E$  is much greater than  $\epsilon$ . Suppose further that  $x$  is measured and found to be in the range indicated by  $q$ . Then if  $p$  denotes this proposition,  $P(p | qh) = 1$ , and  $P(p | \sim qh)$  is of order  $\epsilon/E$ . Hence

$$\frac{P(q | ph)}{P(\sim q | ph)} = O\left(\frac{E}{\epsilon}\right) \frac{P(q | h)}{P(\sim q | h)}.$$

Thus if  $E/\epsilon$  is large and  $q$  is a serious possibility, a single verification may send its probability nearly up to 1. It is an advantage to consider together in this way all hypotheses that would give similar inferences and treat their disjunction as one hypothesis. The data give no information to discriminate between them so long as the data are consequences of all; the posterior probabilities remain in the ratios of the prior probabilities. With this rule, therefore, we can with a few verifications exclude from serious consideration any vaguely stated hypotheses that would require the observed results to be remarkable coincidences; while

unforeseen alternatives whose consequences would agree with those given by hypotheses already included in  $q$ , within the range of verification at any stage, will give no trouble. By the time when any of them is stated explicitly, all hypotheses not implying values of  $x$  within the ranges actually found will have negligible probabilities anyhow, and all that we shall need to do is to separate the disjunction  $q$  as occasion arises. It is therefore desirable as far as possible to state hypotheses in such a form that those with indistinguishable consequences can be treated together; this will avoid mere mathematical complications relating to possibilities that we have no means of testing.

**1.7. THEOREM 11.** *If  $q_1, q_2, \dots, q_n$  are a set of exclusive alternatives on data  $r$ , and if*

$$P(p | q_1 r) = P(p | q_2 r) = \dots = P(p | q_n r),$$

*then each  $= P(p | q_1 \vee q_2 \dots \vee q_n : r)$ .*

For if we denote the disjunction  $q_1 \vee q_2 \dots \vee q_n$  by  $q$ , we have

$$P(pq | r) = P(pq_1 | r) + P(pq_2 | r) + \dots \quad (1)$$

since these alternatives are mutually exclusive; and this

$$= P(p | q_1 r)P(q_1 | r) + \dots \quad (2)$$

The first factors are all equal, and the sum of the second factors is  $P(q | r)$ . Hence

$$P(pq | r) = P(p | q_1 r)P(q | r). \quad (3)$$

But

$$P(pq | r) = P(p | qr)P(q | r), \quad (4)$$

which gives the theorem on comparison with (3).

This leads to the principle that we may call the *suppression of an irrelevant premiss*. If  $q_1 \vee q_2 \dots \vee q_n$  is entailed by  $r$ ,

$$P(p | qr) = P(pq | r) = P(p | r),$$

since  $P(q | pr) = 1$ ; and then each of the expressions  $P(p | q_i r)$  is equal to  $P(p | r)$ . In words, if the probability of a proposition is the same for all the alternative data consistent with one fixed datum, then the probability on the fixed datum alone has the same value.

The interest of this theorem is primarily in relation to what are called 'chances', in a technical sense given by N. R. Campbell and M. S. Bartlett. We have seen that probabilities of propositions in general depend on the data. But cases can be stated, and whether they exist or not must be considered, where the probability is the same over a wide range of data; in such a case we may speak of the information not common to all these data as irrelevant to the probability of the

proposition. Thus above we can say that the propositions  $q_1, \dots, q_n$  are irrelevant to  $p$ , given  $r$ . Further,

$$P(pq_i | r) = P(q_i | r)P(p | q_i r) = P(q_i | r)P(p | r),$$

so that the product formula in such a case is legitimately replaced by the form (2) on p. 27. I shall therefore define a chance as follows: *If  $q_1, q_2, \dots, q_n$  are a set of alternatives, mutually exclusive and exhaustive on data  $r$ , and if the probabilities of  $p$  given any of them and  $r$  are the same, each of these probabilities is called the chance of  $p$  on data  $r$ . It is equal† to  $P(p | r)$ .*

In any case where  $r$  includes the specification of all the parameters in a law, and the results of previous trials are irrelevant to the result of a new trial, the probability of a given result at that trial is the chance on data  $r$ . For the information available just before that trial is made is composed of  $r$  and the results of all previous trials. If we consider the aggregate of all the results that might have been obtained in previous trials, they constitute a set of alternatives such that one of them must occur on data  $r$ , and are exclusive and exhaustive. Given then that the probability of an event at the next trial is the same whatever the results of previous trials, it must be equal to the chance on data  $r$ . It follows that the joint probability on data  $r$  of the results of several trials is the product of their separate chances on data  $r$ . This can easily be proved directly. For if  $p_1, p_2, \dots, p_m$  are the results in order, we have by successive applications of the product formula

$$P(p_1 p_2 \dots p_m | r) = P(p_1 | r)P(p_2 | p_1 r)P(p_3 | p_1 p_2 r) \dots P(p_m | p_1 \dots p_{m-1} r),$$

and by the condition of irrelevance this is equal to

$$P(p_1 | r)P(p_2 | r)P(p_3 | r) \dots P(p_m | r).$$

This is usually taken for granted, but it is just as well to have it proved.

When the probabilities, given the law, are chances, they satisfy the product rule automatically. Hence our proof of the consistency of the principle of inverse probability is complete in all cases where the likelihoods are derived from chances. This covers nearly all the applications in this book.

**THEOREM 12.** *If  $p_1, p_2, \dots, p_m$  and  $q_1, q_2, \dots, q_n$  are two sets of alternatives, each exclusive and exhaustive on data  $r$ , and if*

$$P(p_s q_t | r) = f(p_s)g(q_t)$$

† Bayes and Laplace use both words 'probability' and 'chance', but so far as I know do not specify any distinction between them. There are, however, passages in their writings that suggest that they use the words with their modern senses interchanged.

for all values of  $s$  and  $t$ , where  $f(p_s)$  depends only on  $p_s$  and  $r$ , and  $g(q_t)$  only on  $q_t$  and  $r$ , then

$$P(p_s | r) \propto f(p_s); \quad P(q_t | r) \propto g(q_t).$$

For if we denote the disjunctions of the  $p_s$  and  $q_t$  by  $p$  and  $q$ , we have

$$P(p_s q | r) = \sum_t P(p_s q_t | r) = f(p_s) \sum_t g(q_t), \quad (1)$$

which is proportional to  $f(p_s)$ . But

$$P(p_s q | r) = P(p_s | r) P(q | p_s r) \quad (2)$$

and the last factor is 1 since  $q$  is entailed by  $r$ . Hence

$$P(p_s | r) \propto f(p_s). \quad (3)$$

Similarly,

$$P(q_t | r) \propto g(q_t). \quad (4)$$

We notice that

$$P(pq | r) = \sum_s \sum_t f(p_s) g(q_t) = \sum_s f(p_s) \sum_t g(q_t) \quad (5)$$

and is equal to 1 since  $p$  and  $q$  are both entailed by  $r$ . It is possible to multiply  $f(p_s)$  and  $g(q_t)$  by factors such that both sums will be equal to 1; these factors will be reciprocals; and if this is done, since  $p$  and  $q$  separately are entailed by  $r$ , we shall have

$$P(p_s | r) = f(p_s); \quad P(q_t | r) = g(q_t).$$

Also

$$P(p_s | q_t r) = P(p_s q_t | r) / P(q_t | r) = f(p_s) \quad (6)$$

and  $q_t$  is irrelevant to  $p_s$ .

This theorem is useful in cases where a joint probability distribution breaks up into factors.

**1.8.** Expectation of benefit is taken as a primitive idea in the Bayes-Ramsey theory. In the present one we can define the expectation of a function  $f(x)$  on data  $p$  by the equation

$$E\{f(x) | p\} = \sum f(x) P(x | p)$$

taken over all values of  $x$ . For expectation of benefit, if benefits interfere, there is no great trouble. If  $x$  is, for instance, a monetary gain, we need only distinguish between  $x$  itself, the expectation of which will be  $\sum x P(x | p)$ , and the benefit to us of  $x$ , which is not necessarily proportional to  $x$ . If it is  $f(x)$ , the expectation of benefit will be

$$\sum f(x) P(x | p).$$

The expectations of functions of a variable are often required for our purposes, though we shall not have much more to say about expectation of benefit. But attention must be called at once to the fact that if the expectation of a variable is  $a$ , it does not mean that we expect



the variable to be near  $\alpha$ . Consider the following case. Suppose that we have two boxes  $A$  and  $B$  each containing  $n$  balls. We are to toss a coin; if it comes down heads we shall transfer all the balls from  $A$  to  $B$ ; if tails, all from  $B$  to  $A$ . What is our present expectation of the number of balls in  $A$  after the process? There is a probability  $\frac{1}{2}$  that there will be  $2n$  balls in  $A$ , and a probability  $\frac{1}{2}$  that there will be none. Hence the expectation is  $n$ , which is not a possible value at all. Incorrect results have often been obtained by taking an expectation as a prediction of an actual value; this can be done only if it is also shown that the probabilities of different actual values are closely concentrated about the expectation. It may easily happen that they are concentrated about two or more values, none of which is anywhere near the expectation.

1.9. It may be noticed that the words 'idealism' and 'realism' have not yet been used. I should perhaps explain that their use in everyday speech is different from the philosophical use. In everyday use, realism is thinking that other people are worse than they are; idealism is thinking that they are better than they are. The former is an expression of praise, the latter of disparagement. It is recognized that nobody sees himself as others see him; it follows that everybody knows that everybody else is either a realist or an idealist. In philosophy, realism is the belief that there is an external world, which would still exist if we were not available to make observations, and that the function of scientific method is to find out properties of this world. Idealism is the belief that nothing exists but the mind of the observer or observers and that the external world is merely a mental construct, imagined to give us ourselves a convenient way of describing our experiences. The extreme form of idealism is solipsism, which, for any individual, asserts that only his mind and his sensations exist, other people's minds also being inventions of his own. The methods developed in this book are consistent with some forms of both realism and idealism, but not with solipsism; they contribute nothing to the settlement of the main question of idealism versus realism, but they do lead to the rejection of various special cases of both. I am personally a realist (in the philosophical sense, of course) and shall speak mostly in the language of realism, which is also the language of most people; but if an idealist wishes to translate anything in this book into the language of idealism, I think he will be able to do it. To him I offer the bargain of the Unicorn with Alice: 'If you'll believe in me, I'll believe in you.'

Solipsism is not, as far as I know, actively advocated by anybody (with the possible exception of the behaviourist psychologists). The great difficulty about it is that no two solipsists could agree. If *A* and *B* are solipsists, *A* thinks that he has invented *B* and vice versa. The relation between them is that between Alice and the Red King; but while Alice was willing to believe that she was imagining the King, she found the idea that the King was imagining her quite intolerable. Tweedledum and Tweedledee solved the problem by accepting the King's solution and rejecting Alice's; but every solipsist must have his own separate solipsism, which is flatly contradictory to every other's. Nevertheless, solipsism does contain an important principle, recognized by Karl Pearson, that any person's data consist of his own individual experiences and that his opinions are the result of his own individual thought in relation to those experiences. Any form of realism that denies this is simply false. A hypothesis does not exist till some one person has thought of it; an inference does not exist until one person has made it. We must and do, in fact, begin with the individual. But early in life he recognizes groups of sensations that habitually occur together, and in particular he notices resemblances between those groups that we, as adults, call observations of oneself and other people. When he learns to speak he has already made the observation that some sounds belonging to these groups are habitually associated with other groups of visual or tactile sensations, and has inferred the rule that we should express by saying that particular things and actions are denoted by particular words; and when he himself uses language he has generalized the rule to say that it may be expected to hold for future events.

Thus the use of language depends on the principle that generalization from experience is possible; and this is far from being the only such generalization made in infancy. But if we accept it in one case we have no ground for denying it in another. But a person also observes similarities of appearance and behaviour between himself and other people, and as he himself is associated with a conscious personality, it is a natural generalization to suppose that other people are too. Thus the departure from solipsism is made possible by admitting the possibility of generalization. It is now possible for two people to understand and agree with each other simultaneously, which would be impossible for two solipsists. But we need not say that nothing is to be believed until everybody believes it. The situation is that one person makes an observation or an inference; this is an individual act. If he

reports it to anybody else, the second person must himself make an individual act of acceptance or rejection. All that the first can say is that, from the observed similarities between himself and other people, he would expect the second to accept it. The facts that organized society is possible and that scientific disagreements tend to disappear when the participants exchange their data or when new data accumulate are confirmation of this generalization. Regarded in this way the resemblance between individuals is a legitimate induction, and to take universal agreement as a primary requisite for belief is a superfluous postulate.

Whether one is a realist or an idealist, the problem of inferring future sensations arises, and a theory of induction is needed. Both some realists and some idealists deny this, holding that in some way future sensations can be inferred deductively from some intuitive knowledge of the possible properties of the world or of sensations. If experience plays any part at all it is merely to fill in a few details. This must be rejected under rule 5. I shall use the adjective 'naïve' for any theory, whether realist or idealist, that maintains that inferences beyond the original data are made with certainty, and 'critical' for one that admits that they are not, but nevertheless have validity. Nobody that ever changes his mind through evidence or argument is a naïve realist, though in some discussions it seems to be thought that there is no other kind of realism. It is perfectly possible to believe that we are finding out properties of the world without believing that anything we say is necessarily the last word on the matter.

It should be remarked that some philosophers define 'naïf realism' in some such terms as 'the belief that the external world is something like our perception of it', and argue in its favour. To quote a remark I once heard Russell make, 'I wonder what it feels like to think that.' The succession of two-dimensional impressions that we call visual observations is nothing like the three-dimensional world of science, and I cannot think that such a hypothesis merits serious discussion. The trouble is that many philosophers are as far as most scientists from appreciating the long chain of inference that connects observation with the simplest notions of objects, and many of the problems that take up most attention are either solved at once or are seen to be insoluble when we analyse the process of induction itself.

## II

### DIRECT PROBABILITIES

‘Having thus exposed the far-seeing Mandarin’s inner thoughts, would it be too excessive a labour to penetrate a little deeper into the rich mine of strategy and disclose a specific detail?’

ERNEST BRAMAH, *Kai Lung Unrolls his Mat*

2.0. WE have seen that the principle of inverse probability can be stated in the form

Posterior Probability  $\propto$  Prior Probability  $\times$  Likelihood,

where by the likelihood we understand the probability that the observations should have occurred, given the hypothesis and the previous knowledge. The prior probability of the hypothesis has nothing to do with the observations immediately under discussion, though it may depend on previous observations. Consequently the whole of the information contained in the observations that is relevant to the posterior probabilities of different hypotheses is summed up in the values that they give to the likelihood. In addition, if the observations are to tell us much that we do not know already, the likelihood will have to vary much more between different hypotheses than the prior probability does. Special attention is therefore needed to the discussion of the probabilities of sets of observations given the hypotheses.

Another consideration is that we may be interested in the likelihood as such. There are many problems, such as those of games of chance, where the hypothesis is trusted to such an extent that the amount of observational material that would induce us to modify it would be far larger than will be available in any actual trial. But we may want to predict the result of such a game; or a bridge player may be interested in such a problem as whether, given that he and his partner have nine trumps between them, the remaining four are divided two and two. This is a pure matter of inference from the hypothesis to the probabilities of different events. Such problems have already been treated at great length, and I shall have little to say about them here, beyond indicating their general position in the theory.

In Chapter I we were concerned mainly with the general rules that a consistent theory of induction must follow. They say nothing about what laws actually connect observations; they do provide means of choosing between possible laws, in accordance with their probabilities

given the observations. The laws themselves must be suggested before they can be considered in terms of the rules and the observations. The suggestion is always a matter of imagination or intuition, and no general rules can be given for it. We do not assert that any suggested hypothesis is right, or that it is wrong; it may appear that there are cases where only one is available, but any hypothesis specific enough to give inferences has at least one contradictory, in comparison with which it may be considered. The evaluation of the likelihood requires us to regard the hypotheses as considered propositions, not as asserted propositions; we can give a definite value to  $P(p | q)$  irrespective of whether  $q$  is true or not. This distinction is necessary, because we must be able to consider the consequences of false hypotheses before we can say that they are false.† We get no evidence for a hypothesis by merely working out its consequences and showing that they agree with some observations, because it may happen that a wide range of other hypotheses would agree with those observations equally well. To get evidence for it we must also examine its various contradictories and show that they do not fit the observations. This elementary principle is often overlooked in alleged scientific work, which proceeds by stating a hypothesis, quoting masses of results of observation that might be expected on that hypothesis and possibly on several contradictory ones, ignoring all that would not be expected on it, but might be expected on some alternative, and claiming that the observations support the hypothesis. Most of the current presentations of the theory of relativity (the essentials of which are supported by observation) are of this type; so are those of the theory of continental drift (the hypotheses of which are contradicted by every other check that has been applied). So long as alternatives are not examined and compared with the whole of the relevant data, a hypothesis can never be more than a considered one.

In general the probability of an empirical proposition is subject to some considered hypothesis, which usually involves a number of quantitative parameters. Besides this, the general principles of the theory and of pure mathematics will be part of the data. It is convenient to have a summary notation for the set of propositions accepted throughout an investigation; I shall use  $H$  to denote it.  $H$  will include the specification of the conditions of an observation.  $\theta$  will often be used to denote the observational data.

† This is the reason for rejecting the *Principia* definition of implication, which leads to the proposition, 'If  $q$  is false, then  $q$  implies  $p$ .' Thus any observational result  $p$  could be regarded as confirming a false hypothesis  $q$ . In terms of entailment the corresponding proposition, 'If  $q$  is false,  $q$  entails  $p$ ', does not hold irrespective of  $p$ .

**2.1. Sampling.** Suppose that we have a population, composed of members of two types  $\phi$  and  $\sim\phi$ , in known numbers. A sample of given number is drawn in such a way that any set of that number in the population is equally likely to be taken. What, on these data, is the probability that the numbers of the two types will have a given pair of values?

Let  $r$  and  $s$  be the numbers of types  $\phi$  and  $\sim\phi$  in the population,  $l$  and  $m$  those in the sample. The number of possible samples, subject to the conditions, is the number of ways of choosing  $l+m$  things from  $r+s$ , which we denote by  ${}^{r+s}C_{l+m}$ . The number of them that will have precisely  $l$  things of type  $\phi$  and  $m$  of type  $\sim\phi$  is  ${}^rC_l {}^sC_m$ . Now on data  $H$  any two particular samples are exclusive alternatives and are equally probable; and some sample of total number  $l+m$  must occur. Hence the probability that any particular sample will occur is  $1/{}^{r+s}C_{l+m}$ ; and the probability that the actual numbers will be  $l$  and  $m$  is obtained, by the addition rule, by multiplying this by the total number of samples with these numbers. Hence

$$P(l, m | H) = {}^rC_l {}^sC_m / {}^{r+s}C_{l+m}. \quad (1)$$

It is an easy algebraic exercise to verify that the sum of all these expressions for different values of  $l, l+m$  remaining the same, is 1.

Explicit statement of the data  $H$  is desirable because it may be true in some cases that all samples are possible but not equally probable. In such cases the application of the rule may lead to results that are seriously wrong. To obtain a genuine random sample involves indeed a difficult technique. Yule and Kendall give examples of the dangers of supposing that a sample taken without any particular thought is a random sample. They are all rather more complicated than this problem. But the following would illustrate the point. Suppose that we want to know the general opinion of British adults on a political question. The most thorough method would be a referendum to the entire electorate. But a newspaper may attempt to find it by means of a vote among its readers. These will include many regular subscribers, and also many casual purchasers. It is possible that on a given day any individual might obtain the paper—even if it was only because all the others were sold out. Thus all the conditions in  $H$  are satisfied, except that of randomness; because on the day when the voting-papers are issued there is not an equal chance of a regular subscriber and an occasional purchaser obtaining that particular number of the paper. The tendency of such a vote would therefore be to give an excess chance

of a sample containing a disproportionately high number of regular subscribers, who would presumably be more in sympathy with the general policy of the paper than the bulk of the population.

**2.11.** Another type of sampling, which is extensively discussed in the literature, is known as sampling with replacement. In this case every member, after being examined, is replaced before the next draw. At each stage every member, whether previously examined or not, is taken to be equally likely to be drawn at any particular draw. This is not true in simple sampling, because a member already examined cannot be drawn at the next draw. If  $r$  and  $s$  as before are the numbers of the types in the population, the chance at any draw of a member of the first type being drawn, given the results of all the previous draws, will always be  $r/(r+s)$ , and that of one of the second type  $s/(r+s)$ . This problem is a specimen of the cases where the probabilities reduce to chances.

Many other actual cases are chances or approximate to them. Thus the probabilities that a coin will throw a head, or a die a 6, appear to be chances, as far as we can tell at present. This may not be strictly true, however, since either, if thrown a sufficient number of times, would in general wear unevenly, and the probability of a head or a six on the next throw, given all previous throws, would depend partly on the amount of this wear, which could be estimated by considering the previous throws. Thus it would not be a chance. The existence of chances in these cases would not assert that the chance of a head is  $\frac{1}{2}$  or that of a six  $\frac{1}{6}$ ; the latter indeed seems to be untrue, though it is near enough for most practical purposes.

If the chance of an event of the first type (which we may now call a success) is  $x$ , and that of one of the second, which we shall call a failure, is  $1-x = y$ , then the joint probability that  $l+m$  trials will give just  $l$  successes and  $m$  failures, in any prescribed order, is  $x^l y^m$ . But there will be  ${}^{l+m}C_l$  ways of assigning the  $l$  successes to possible positions in the series, and these are all equally probable. Hence in this case

$$P(l, m | H) = \frac{(l+m)!}{l! m!} x^l y^m, \quad (2)$$

which is a typical term in the binomial expression for  $(x+y)^{l+m}$ . Hence this law is usually known as the *binomial distribution*. In the case of sampling with replacement it becomes

$$P(l, m | H) = \frac{(l+m)!}{l! m!} \left( \frac{r}{r+s} \right)^l \left( \frac{s}{r+s} \right)^m. \quad (3)$$

It is easy to verify that with either type of sampling the most probable value of  $l$  is within one unit of  $r(l+m)/(r+s)$ , so that the ratio of the types in the sample is approximately the ratio in the population sampled. This may be expressed by saying that in the conditions of random sampling or sampling with replacement the most probable sample is a fair one. It can also be shown easily that if we consider in succession larger and larger populations sampled, the size of the sample always remaining the same, but  $r$  and  $s$  tending to infinity in such a way that  $r/s$  tends to a fixed value  $x/y$ , the formula for simple sampling tends to the binomial one. What this means is that if the population is sufficiently large compared with the sample, the extraction of the sample makes a negligible difference to the probability at the next trial, which can therefore be regarded as a chance with sufficient accuracy.

2.12. Consider now what happens to the binomial law when  $l$  and  $m$  are large and  $x$  fixed. Let us put

$$\frac{1}{f(l)} = \frac{(l+m)!}{l!m!} x^l y^m, \quad (4)$$

$$l+m = n; \quad l = nx + n^{1/2}\alpha; \quad m = ny - n^{1/2}\alpha, \quad (5)$$

and suppose that  $\alpha$  is not large. Then

$$\log f(l) = \log l! + \log m! - \log n! - l \log x - m \log y. \quad (6)$$

Now we have Stirling's formula†

$$\log n! = (n + \frac{1}{2}) \log n - n + \frac{1}{2} \log 2\pi + \frac{1}{12n} - O\left(\frac{1}{n^3}\right). \quad (7)$$

Substituting and neglecting terms of order  $1/l$ ,  $1/m$ , we have

$$\log f(l) = \frac{1}{2} \log \frac{2\pi lm}{n} + l \log \frac{l}{nx} + m \log \frac{m}{ny}. \quad (8)$$

† The closeness of Stirling's approximation, even if  $1/12n$  is neglected, is remarkable. Thus for  $n = 1$  and  $2$  it gives

$$1! = 0.9221; \quad 2! = 1.9190;$$

while if the term in  $1/12n$  is kept it gives

$$1! = 1.0022; \quad 2! = 2.0006.$$

Considered as approximations on the hypothesis that  $1$  and  $2$  are large numbers they are very creditable. The use of the logarithmic series may lead to larger errors.

Proofs of the formula and of other properties of the factorial function, not restricted to integral argument, are given in H. and B. S. Jeffreys, *Methods of Mathematical Physics*, Chapter 15.



Now substituting for  $l$  and  $m$ , and expanding the logarithms to order  $\alpha^2$  we have

$$\log f(l) = \frac{1}{2} \log(2\pi nxy) + \frac{\alpha^2}{2xy} + O(\alpha^3 l^{-1/2}, \alpha^3 m^{-1/2}), \quad (9)$$

$$\frac{1}{f(l)} \doteq \frac{1}{(2\pi nxy)^{1/2}} \exp\left\{-\frac{(l-nx)^2}{2nxy}\right\}. \quad (10)$$

This form is due to De Moivre.† From inspection of the terms neglected we see that this will be a good approximation if  $l$  and  $m$  are large and  $\alpha$  not large compared with  $l^{1/6}$  or  $m^{1/6}$ . Also if  $nxy$  is large the chance varies little between consecutive values of  $l$ , and the sum over a range of values may be closely replaced by an integral, which will be valid as an approximation till  $l-nx$  is more than a few times  $(nxy)^{1/2}$ . But the integrand falls off with  $l-nx$  so rapidly that the integral over the range where (10) is valid is practically 1, and therefore includes nearly all the chance. But the whole probability of all values of  $l$  is 1. It follows that nearly the whole probability of values of  $l$  is concentrated in a range such that (10) is a good approximation to (4).

It follows further that if we choose any two positive numbers  $\beta$  and  $\gamma$ , and consider the probability that  $l$  will lie between  $n(x+\beta)$  and  $n(x-\gamma)$ , it will be approximately

$$\left(\frac{n}{2\pi xy}\right)^{1/2} \int_{-\gamma}^{\beta} \exp\left(-\frac{nz^2}{2xy}\right) dz, \quad (11)$$

which, if  $\beta$  and  $\gamma$  remain fixed, will tend to 1 as  $n$  tends to infinity. That is, the probability that  $(l-nx)/n$  will lie within any specified limits, however close, provided that they are of opposite signs, will tend to certainty.

**2.13.** This theorem was given by James Bernoulli in the *Ars Conjectandi* (1713). It is sometimes known as the law of averages or the law of large numbers. It is an important theorem, though it has often been misinterpreted. We must notice that it does not prove that the ratio  $l/n$  will tend to limit  $x$  when  $n$  tends to infinity. It proves that, subject to the probability at every trial remaining the same, however many trials we make, and whatever the results of previous trials, *we may reasonably expect* that  $l/n-x$  will lie within any specified range about 0 for any *particular* value of  $n$  greater than some assignable one depending on this range. The larger  $n$  is, the more closely will this probability approach to certainty, tending to 1 in the limit. The

† *Miscellanea Analytica*, 1733.

existence of a limit for  $l/n$  would require that there shall be a series of positive numbers  $\alpha_n$ , depending on  $n$  and tending to 0 as  $n \rightarrow \infty$ , such that, for all values of  $n$  greater than some specified  $n_0$ ,  $l/n - x$  lies between  $\pm \alpha_n$ . But it cannot be proved mathematically that such series always exist when the sampling is random. Indeed we can produce possible results of random sampling where they do not exist. Suppose that  $x = \frac{1}{2}$ . It is essential to the notion of randomness that the results of previous trials are irrelevant to the next. Consequently we can never say at any definite stage that a particular result is out of the question. Thus if we enter 1 for each success and 0 for each failure such series as the following could arise:

100110010100100111010...,  
 100100100100100100100...,  
 00000000000000000000...,  
 11111111111111111111...,  
 10110000111111110000000000....

The first series was obtained by tossing a coin. The others were systematically designed; but it is impossible to say logically at any stage that the conditions of the problem forbid the alternative chosen. They are all possible results of random sampling consistent with a chance  $\frac{1}{2}$ . But the second would give limit  $\frac{1}{3}$ ; the third and fourth limits 0 and 1; the fifth would give no limit at all, the ratio  $l/n$  oscillating between  $\frac{1}{3}$  and  $\frac{2}{3}$ . (The rule adopted for this is that the number of zeros or units in each block is equal to the whole number of figures before the beginning of the block.) An infinite number of series could be chosen that would all be possible results of random selection, assuming an infinite number of random selections possible at all, and giving either a limit different from  $\frac{1}{2}$  or no limit.

It was proved by Wrinch and me,<sup>†</sup> and another version of the proof is given by M. S. Bartlett,<sup>‡</sup> that if we take a *fixed*  $\alpha$  independent of  $n$ ,  $n_0$  can always be chosen so that the probability that there will be no deviation numerically greater than  $\alpha$ , for any  $n$  greater than  $n_0$ , is as near 1 as we like. But since the required  $n_0$  tends to infinity as  $\alpha$  tends to 0, we have the phenomenon of convergence with infinite slowness that led to the introduction of the notion of uniform convergence. It is necessary, to prove the convergence of the series, that  $\alpha_n$  shall tend to 0; it must not be independent of  $n$ , otherwise the ratio might oscillate finitely for ever.

<sup>†</sup> *Phil. Mag.* 38, 1919, 718-19.

<sup>‡</sup> *Proc. Roy. Soc. A*, 141, 1933, 520-1.

Before considering this further we need a pair of bounds for the incomplete factorial function,

$$I = \int_x^{\infty} u^n e^{-tu} du, \quad (1)$$

where  $x$  is large. Then

$$I > x^n \int_x^{\infty} e^{-tu} du = x^n e^{-tx}/t. \quad (2)$$

Also, if  $u = x + v$ ,

$$u/x < \exp v/x,$$

$$I < x^n e^{-tx} \int_0^{\infty} \exp\left(-t + \frac{n}{x}\right)v dv = \frac{x^n e^{-tx}}{t - n/x}. \quad (3)$$

Hence, if  $x/n$  is large,

$$I = \frac{x^n e^{-tx}}{t} \left[ 1 + O\left(\frac{n}{x}\right) \right]. \quad (4)$$

Now let  $P(n)$  be the chance of a ratio in  $n$  trials outside the range  $x \pm \alpha$ . This is asymptotically

$$\begin{aligned} P(n) &\sim 2 \int_{\alpha}^{\infty} \left\{ \frac{n}{2\pi x(1-x)} \right\}^{1/2} \exp\left[ -\frac{n\alpha^2}{2x(1-x)} \right] d\alpha \\ &= \left( \frac{2x(1-x)}{\pi n} \right)^{1/2} \frac{1}{\alpha} \exp\left\{ -\frac{n\alpha^2}{2x(1-x)} \right\} \{1 + O(n^{-1/2})\} \end{aligned} \quad (5)$$

by putting  $\alpha^2 = u$  and applying (4).

Now take

$$\alpha_n = n^{-1/4}. \quad (6)$$

The total chance that there will be a deviation greater than  $\alpha_n$ , for some  $n$  greater than  $n_0$ , is less than the sum of the chances for the separate  $n$ , since the alternatives are not exclusive. Hence this chance

$$\begin{aligned} Q(n_0) &< \sum_{n=n_0}^{\infty} \left\{ \frac{2x(1-x)}{\pi} \right\}^{1/2} n^{-1/4} \exp\left[ -\frac{n^{1/2}}{2x(1-x)} \right] \\ &\sim \int_{n_0}^{\infty} \left\{ \frac{2x(1-x)}{\pi} \right\}^{1/2} n^{-1/4} \exp\left[ -\frac{n^{1/2}}{2x(1-x)} \right] dn. \end{aligned} \quad (7)$$

Put

$$n = u^2;$$

then

$$\begin{aligned} Q(n_0) &< \int_{\sqrt{n_0}}^{\infty} 2 \left\{ \frac{2x(1-x)}{\pi} \right\}^{1/2} u^{1/2} \exp\left[ -\frac{u}{2x(1-x)} \right] du \\ &< \frac{2\{2x(1-x)\}^{3/2}}{\pi^{1/2}} n_0^{1/4} \exp\left[ -\frac{n_0^{1/2}}{2x(1-x)} \right] \end{aligned} \quad (8)$$

with a correcting term small compared with the first for large  $n_0$ . Hence  $Q(n_0)$  does tend to zero as  $n_0$  tends to infinity, and we have the result that  $n_0$  can be fixed so that the total chance of deviations greater than  $n^{-1/4}$  for all  $n$  greater than  $n_0$  is as small as we please; and if all deviations are less than  $n^{-1/4}$  the series converges. Hence it may be expected, with an arbitrarily close approach to certainty, that subject to the conditions of random sampling the ratio in the series will tend to  $x$  as a limit.†

This, however, is still a probability theorem and not a mathematically proved one; the mathematical theorem, that the limit must exist in any case, is false because exceptions that are possible in the conditions of random sampling can be stated.

The situation is that the proposition that the ratio does not tend to limit  $x$  has probability 0 in the conditions stated. This, however, does not entail that it will tend to this limit. We have seen (1) that series such that the ratio does not tend to limit  $x$  are possible in the conditions of the problem, (2) that though a proposition impossible on the data must have probability 0 on those data, the converse is not true; a proposition can have probability 0 and yet be possible in much simpler cases than this, if we maintain Axiom 5, that probabilities on given data form a set of not higher ordinal type than the continuum. If a magnitude, limited to a continuous set of positive values, is less than any assignable positive quantity, then it is 0. But this is not a contradiction because the converse of Theorem 2 is false. We need only distinguish between propositions logically contradicted by the data, in which case the impossibility can be proved by the methods of deductive logic, and propositions possible on the data but whose probability is zero, such as that a quantity with a uniform distribution of its probability between 0 and 1 is exactly  $\frac{1}{2}$ .

The result is not of much practical importance; we never have to count an infinite series empirically given, and though we might like to make inferences about such series we must remember the condition required by Bernoulli's theorem, that no number of trials, however large, can possibly tell us anything about their immediate successor that we did not know at the outset. It seems that in physical conditions something analogous to the wear of a coin would always violate this condition. Consequently it appears that the problem could never arise. Further, there is a logical difficulty about whether the limit of a ratio

† Another proof is given by F. P. Cantelli, *Rend. d. circ. matem., Palermo*, **41**, 1916, 191–201; *Rend. d. R. Acad. d. Lincei*, **26**, 1917, 39–45. See E. C. Fieller, *J. R. Stat. Soc.* **99**, 1936, 717.

in a random series has any meaning at all. In the infinite series considered in mathematics a law connecting the terms is always given, and the sum of any number of terms can be calculated by simply following rules stated at the start. If no such law is given, which is the essence of a random process, there is no means of calculation. The difficulty is associated with what is called the Multiplicative Axiom; this asserts that such a rule always exists, but it has not been proved from the other axioms of mathematical logic, though it has recently been proved by Gödel to be consistent with them. Littlewood† remarks, 'Reflection makes the intuition of its truth doubtful, analysing it into prejudices derived from the finite case, and short of intuition there seems to be nothing in its favour.' The physical difficulty may arise in a finite number of trials, so that there is no objection to supposing that it may arise in any case even if the Multiplicative Axiom is true. In fact I should say that the notion of chance is never more than a considered hypothesis that we are at full liberty to reject. Its usefulness is not that chances ever exist, but that it is sufficiently precisely stated to lead to inferences definite enough to be tested, and when it is found wrong we shall in the process find out how much it is wrong.

**2.14.** We can use the actual formula 2.12 (10) to obtain an approximation to the formula for simple sampling when  $l$ ,  $m$ ,  $r-l$ , and  $s-m$  are all large. Consider the expression

$$F = {}^r C_l x^l y^{r-l} \times {}^s C_m x^m y^{s-m}, \quad (1)$$

where  $x$  and  $y$  are two arbitrary numbers subject to  $x+y=1$ .  $r$ ,  $s$ , and  $l+m$  are fixed. Choose  $x$  so that the maxima of the two expressions multiplied are at the same value of  $l$ , and call this value  $l_0$  and the corresponding value of  $m$ ,  $m_0$ . Then

$$l_0 = rx; \quad r-l_0 = ry; \quad m_0 = sx; \quad s-m_0 = sy; \quad (2)$$

whence

$$(r+s)x = l_0 + m_0 = l + m. \quad (3)$$

Then, by 2.12 (10),

$$\begin{aligned} F &\doteq (2\pi rxy)^{-1/2} \exp\left\{-\frac{(l-l_0)^2}{2rxy}\right\} (2\pi sxy)^{-1/2} \exp\left\{-\frac{(m-m_0)^2}{2sxy}\right\} \\ &= (2\pi xy)^{-1} (rs)^{-1/2} \exp\left\{-\frac{(l-l_0)^2(r+s)}{2rsxy}\right\}. \end{aligned} \quad (4)$$

$$\text{Also} \quad G = {}^{r+s} C_{l+m} x^{l+m} y^{r+s-l-m} \doteq \{2\pi(r+s)xy\}^{-1/2}. \quad (5)$$

† *Elements of the Theory of Real Functions*, 1926, p. 25.

Hence by division

$$P(l, m | H) = \frac{{}^r C_l {}^s C_m}{{}^{r+s} C_{l+m}} \doteq \left( \frac{r+s}{2\pi r s x y} \right)^{1/2} \exp \left\{ -\frac{(l-l_0)^2 (r+s)}{2 r s x y} \right\}. \quad (6)$$

$$\text{But} \quad (r+s)^2 x y = (l+m)(r+s-l-m), \quad (7)$$

whence

$$P(l, m | H) \doteq \left\{ \frac{(r+s)^3}{2\pi r s (l+m)(r+s-l-m)} \right\}^{1/2} \exp \left\{ -\frac{(l-l_0)^2 (r+s)^3}{2 r s (l+m)(r+s-l-m)} \right\}, \quad (8)$$

$$\text{where} \quad l_0 = \frac{r(l+m)}{r+s}. \quad (9)$$

Comparing this with 2.12 (10) we see that it is of similar form, and the same considerations about the treatment of the tail will apply. If  $r$  and  $s$  are very large compared with  $l$  and  $m$ , we can write

$$r = (r+s)p; \quad s = (r+s)q, \quad (10)$$

$p$  and  $q$  now corresponding to the  $x$  and  $y$  of the binomial law; and the result approximates to

$$\left\{ \frac{1}{2\pi(l+m)pq} \right\}^{1/2} \exp \left[ -\frac{\{l-p(l+m)\}^2}{2(l+m)pq} \right], \quad (11)$$

which is equivalent to 2.12 (10). In this form we see that the probabilities of different compositions of the sample depend only on the sample and on the ratio of the type numbers in the population sampled; provided that the population is large compared with the sample, further information about its size is practically irrelevant. But in general, on account of the factor  $(r+s)/(r+s-l-m)$  in the exponent, the probability will be somewhat more closely concentrated about the maximum than for the corresponding binomial. This represents the effect of the withdrawal of the first parts of the sample on the probabilities of the later parts, which will have a tendency to correct any departure from fairness in the earlier ones.

**2.15. Multiple sampling and the multinomial law.** These are straightforward extensions of the laws for simple sampling and the binomial law. In the first case, the population consists of  $p$  different types instead of two, the numbers being  $r_1, r_2, \dots, r_p$ ; the corresponding numbers in the sample are  $n_1, n_2, \dots, n_p$  with a prescribed total. It is supposed as before that all possible samples of the given total number are equally probable. The result is

$$P(n_1, n_2, \dots, n_p | H) = {}^{r_1} C_{n_1} {}^{r_2} C_{n_2} \dots {}^{r_p} C_{n_p} / {}^{\Sigma r} C_{\Sigma n}. \quad (1)$$

In the second case, the chances of the respective types occurring at any trial are  $x_1, x_2, \dots, x_p$  (their total being 1) and the number of trials  $\sum n$  is prescribed. The result is

$$P(n_1, n_2, \dots, n_p | H) = \frac{(\sum n)!}{n_1! n_2! \dots n_p!} x_1^{n_1} x_2^{n_2} \dots x_p^{n_p}. \quad (2)$$

It is easy to verify in (1) that the most probable set of values of the  $n$ 's are nearly in the ratios of the  $r$ 's, and in (2) that the most probable set are nearly in the ratios of the  $x$ 's. Consequently we may in both cases speak of the expected or calculated values; if  $N$  is the prescribed total number of the sample, the expected  $n_1$  for multiple sampling will be  $Nr_1/\sum r$ , and the expected  $n_1$  for the multinomial will be  $Nx_1$ . The probability will, however, in both cases be spread over a range about the most probable values, and we shall need to attend later to the question of how great a departure from the most probable values, on the hypothesis we are considering, can be tolerated before we can say that there is evidence against the hypothesis.

**2.16. The Poisson law.**† We have seen that the use of Stirling's formula in the approximation used for the binomial law involves the neglect of terms of order  $1/l$  and  $1/m$ , while the result shows that there is a considerable probability of departures of  $l$  from  $nx$  of amounts of order  $(nxy)^{1/2}$ . If then  $(nxy)^{1/2} > nx$ , the result shows that  $l = 0$  is a very probable value, and the approximation must fail. But if  $n$  is large, this condition implies that  $x$  is small enough for  $nx$  to be less than 1. Special attention is therefore needed to cases where  $n$  is large but  $nx$  moderate. We take the binomial law in the form

$$P(l | H) = \frac{n!}{l!(n-l)!} x^l (1-x)^{n-l}. \quad (1)$$

$$\text{Now} \quad \log\{n!/(n-l)!\} = l \log n + O(l^2/n). \quad (2)$$

$$\text{Also, since } x \text{ is small,} \quad (1-x)^{n-l} = e^{-(n-l)x} \quad (3)$$

nearly; whence, so long as  $l^2/n$  and  $lx$  are small,

$$P(l | H) = \frac{(nx)^l}{l!} e^{-nx}. \quad (4)$$

The sum of this for all values of  $l$  is unity, the terms being  $e^{-nx}$  times the terms of the expansion of  $e^{nx}$ . The formula is the limit of the binomial when  $n$  tends to infinity and  $x$  to 0, but  $nx$  to a definite value. If  $nx^2$  is small but  $nx$  large, both approximations to the binomial are valid.

The condition for the Poisson law is that there shall be a small chance

† S. D. Poisson, *Recherches sur la probabilité des jugements*, 1837, pp. 205-7.

of an event in any one trial, but there are so many trials that there is an appreciable probability that the event will occur in some of them. One of the best-known cases is the study of von Bortkiewicz on the number of men killed by the kick of a horse in certain Prussian army corps in twenty years. The unit being one army corps for one year, the data for fourteen corps for twenty years gave the following summary.†

<i>Number of deaths</i>	<i>Number of units</i>	<i>Expected</i>
0	144	139.0
1	91	97.3
2	32	34.1
3	11	8.0
4	2	1.4
5 and more	0	0.2

The analysis here would be that the chance of any one man being killed by a horse in a year is small, but the number of men in an army corps is such that the chance that there will be one man killed in an entire corps is appreciable. The probabilities that there will be 0, 1, 2, ... men killed in a corps in a year are therefore given by the Poisson rule; and then by the multinomial rule, in a sample of 280 units, we should expect the observed numbers to be in approximately the ratios of these probabilities. The column headed 'expected' gives the expectations on the hypothesis that  $nx = 0.70$ . They have been recalculated, the calculated values as quoted having been derived from several Poisson laws superposed.

Another instance is radioactive disintegration. The chance of a particular atom of a radioactive element breaking up in a given interval may be very small; but a specimen of the substance may contain something of the order of  $10^{20}$  atoms, and the chance that some of them may break up is appreciable. The following table, due to Rutherford and Geiger,‡ gives the observed and expected numbers of intervals of  $\frac{1}{2}$  minute when 0, 1, 2, ...  $\alpha$ -particles were ejected by a specimen.

<i>Number</i>	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Obs.	57	203	383	525	532	408	273	139	45	27	10	4	0	1	1
Exp.	54	211	407	525	508	393	254	140	68	29	11	4	1	0	0
O-E	+3	-8	-24	0	+24	+15	+19	-1	-23	-2	-1	0	-1	+1	+1

$nx$  is taken as the total number of particles divided by the total number of intervals =  $10097/2608 = 3.87$ . It is clear that the Poisson law agrees with the observed variation within about one-twentieth of its range; a closer check will be given later.

† von Bortkiewicz, *Das Gesetz d. kleinen Zahlen*, 1898. Quoted by Keynes, p. 402.

‡ Rutherford, H. Geiger, and H. Bateman, *Phil. Mag.* 20, 1910, 698-707.



The Aitken dust-counter provides an example from meteorology.† The problem is to estimate the number of dust nuclei in the air. A known volume of air is admitted into a chamber containing moisture and filtered air, and is then made to expand. This causes condensation to take place on the nuclei. The drops in a small volume fall on to a stage and are counted. Here the large number is the number of nuclei in the chamber, the small chance is the chance that any particular one will be within the small volume at the moment of sampling. Scrase gives the following values.

Number	0	1	2	3	4	5	6	7	8
Obs.	23	56	88	95	73	40	17	5	3
Exp.	25	65	88	82	61	38	21	10	4
O-E	-2	-9	0	+13	+12	+2	-4	-5	-1

The data are not homogeneous, the observations having been made on twenty different days;  $nx$  was estimated separately for each and these separate expectations were calculated and added. It appears that the method gives a fair representation of the observed counts, though there are signs of a systematic departure. Scrase suggests that in some cases zero counts may have been wrongly rejected under the impression that the instrument was not working. This would lead to an overestimate of  $nx$  on some days, therefore to an overestimate of the expectations for large numbers, and therefore to negative residuals at the right of the table. Mr. Diananda points out that the observed counts agree quite well with  $nx = 2.925$ .

**2.2. The normal law of error.** Let us suppose that a quantity that we are trying to measure is equal to  $\lambda$ , but that there are various possible disturbances,  $n$  in number, each of which in any particular case has equal chances  $\frac{1}{2}$  of producing alterations  $\pm\epsilon$  in the actual measure; the sign of the contribution from each is independent of those of the others. This is a case of the binomial law. If  $l$  of the components in an individual observation are positive and the remaining  $n-l$  negative, the measured value will be

$$x = \lambda + l\epsilon - (n-l)\epsilon = \lambda + (2l-n)\epsilon. \quad (1)$$

The possible measured values will then differ from  $\lambda - n\epsilon$  by even multiples of  $\epsilon$ . We suppose  $n$  large. Then the probabilities of different values of  $l$  are distributed according to the law obtained by putting  $x = y = \frac{1}{2}$  in 2.12 (10), namely,

$$P(l|H) = \left(\frac{2}{\pi n}\right)^{1/2} \exp\left\{-\frac{2}{n}(l-\frac{1}{2}n)^2\right\}, \quad (2)$$

† John Aitken, *Proc. Roy. Soc. Edin.* **16**, 1888, 135-72; F. J. Scrase, *Q.J.R. Met. Soc.* **61**, 1935, 368-78.

and the probability that  $l$  will be equal to  $l_1, l_2 (> l_1)$ , or some intermediate value will be

$$P(l_2 \geq l \geq l_1 | H) = \sum_{l=l_1}^{l_2} \left( \frac{2}{\pi n} \right)^{1/2} \exp \left\{ -\frac{2(l - \frac{1}{2}n)^2}{n} \right\}. \quad (3)$$

But this is the probability that the measure  $x$  will be in the range from  $\lambda + (2l_1 - n)\epsilon$  to  $\lambda + (2l_2 - n)\epsilon$ , inclusive. If, then, we consider a range  $x_1$  to  $x_2$ , long enough to include many possible values of  $l$ , we can replace the sum by an integral, write

$$l - \frac{1}{2}n = (x - \lambda)/2\epsilon, \quad (4)$$

and 
$$P(x_2 \geq x \geq x_1 | H) = \sum_{x=x_1}^{x_2} \left( \frac{2}{\pi n} \right)^{1/2} \exp \left\{ -\frac{(x - \lambda)^2}{2n\epsilon^2} \right\}. \quad (5)$$

This range will contain  $(x_2 - x_1)/2\epsilon + 1$  admissible values of  $x$ . Now suppose that  $x_2 - x_1$ , which is much larger than  $\epsilon$ , is also much less than  $\epsilon\sqrt{n}$ . The sum will then approximate to

$$\int_{x_1}^{x_2} \left( \frac{2}{\pi n} \right)^{1/2} \exp \left\{ -\frac{(x - \lambda)^2}{2n\epsilon^2} \right\} \frac{dx}{2\epsilon}. \quad (6)$$

Now let  $n$  be very large and  $\epsilon$  very small, in such a way that  $\epsilon\sqrt{n}$  is finite. The possible values of  $x$  will then become indefinitely closely packed, and if we now consider a small range from  $x_1$  to  $x_1 + \delta x$ , the chance that  $x$  lies within it will approximate to

$$P(x_1 < x < x_1 + \delta x | H) = \frac{1}{(2\pi n)^{1/2}\epsilon} \exp \left\{ -\frac{(x_1 - \lambda)^2}{2n\epsilon^2} \right\} \delta x. \quad (7)$$

This is an instance of the normal law, which we can write in its general form

$$P(x_1 < x < x_1 + dx | H) = \frac{1}{\sqrt{(2\pi)\sigma}} \exp \left\{ -\frac{(x_1 - \lambda)^2}{2\sigma^2} \right\} dx, \quad (8)$$

or, more briefly,

$$P(dx | H) = \frac{1}{\sqrt{(2\pi)\sigma}} \exp \left\{ -\frac{(x - \lambda)^2}{2\sigma^2} \right\} dx,$$

in the sense that when  $dx$  tends to zero the ratio of the two sides tends to 1. In practice we are always concerned with finite ranges, so that strictly we always require the integrals of these expressions over some finite range, and the transition from  $\delta x$  to  $dx$  involves only a step that we shall always undo before we make any use of the results.

It will be noticed that whereas we started with three parameters  $\lambda$ ,  $n$ , and  $\epsilon$ , in the result we are left with two,  $\lambda$  and  $\epsilon\sqrt{n}$ , the latter being replaced by  $\sigma$ . This is similar to what happens in sampling, where the

size of the population sampled becomes irrelevant when it is large. The form of the normal law, in application to errors, seems to have been given first by Laplace in 1783, though it is usually attributed to Gauss.†

The law can also be written

$$P(x_1 < x < x_1 + dx | H) = \frac{h}{\sqrt{\pi}} \exp\{-h^2(x-\lambda)^2\} dx, \quad (9)$$

where  $2h^2\sigma^2 = 1$ . (10)

$\sigma$  is usually called the *standard error*, but sometimes the mean square error or simply the mean error.  $h$  is called the precision constant. If we introduce the error function

$$\operatorname{erf} x = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt, \quad (11)$$

the probability that  $x$  will be less than  $x_1$  is  $\frac{1}{2}\{1 + \operatorname{erf} h(x_1 - \lambda)\}$ . Tables of the probability that  $x - \lambda$  will be less than given multiples of  $\sigma$  are given by Sheppard and by later writers. The error function, which has other applications in heat conduction and diffusion, is tabulated by Milne-Thomson and Comrie. In statistical applications (8) is more convenient than (11), since  $\sigma$  usually arises more directly than  $h$ . The curve  $y \propto \exp\{-(x-\lambda)^2/2\sigma^2\}$  has inflexions at  $\lambda \pm \sigma$ . There is a probability 0.683 that an observation will lie between  $\lambda \pm \sigma$ . There is a probability  $\frac{1}{2}$  that it will lie between  $\lambda \pm 0.6745\sigma$ . In this sense  $0.6745\sigma$  is often called the *probable error*, and is the uncertainty usually quoted in astronomical and physical works. This practice would be better abandoned. In applying any significance test or the  $\chi^2$  or  $t$  rules what arises is  $\sigma$ , and if uncertainties are given in terms of the probable error, the multiplication must first be undone, with unnecessary trouble and some loss of accuracy due to accumulation of rounding-off errors.

The conditions contemplated in the normal law of error have often a rough justification. In many cases we have adequate reason to suppose that the quantity we are trying to measure has a 'true value', though we must reserve a further discussion of what that can mean in relation to our general theory. But several minor disturbances may affect any individual measure, such as wandering of the observer's attention, the fact that he must round off his measures to the nearest multiple or tenth of the scale interval, disturbance of the apparatus through vibration of the ground or wind, and so on. These can often be regarded as independent. They are not in general capable of producing only two

† Pearson, *Biometrika*, 13, 1920, 25.

equal and opposite values<sup>\*</sup> of the disturbance; most of them are capable of a continuous range of values, and in general there is not much reason to suppose that these are equally spread for all the disturbances. The general application of the above argument must therefore be mistrusted. It can be regarded only as an indication that there may be cases where the chance of error is distributed according to the normal law, which sums up the whole information with regard to the possible variation in two parameters  $\lambda$  and  $\sigma$ .  $\lambda$  is also often called the population mean and  $\sigma$  the population standard deviation. The latter term is rather cumbrous, and if the word 'population' is omitted it is liable to be confused with the standard deviation of a given finite set of observations, which is not the same thing.

Where we are dealing with a law of the form

$$P(dx | H) = f\left(\frac{x-\lambda}{\sigma}\right) \frac{dx}{\sigma},$$

of which the normal law is an instance, we may speak of  $\lambda$  as the location parameter and  $\sigma$  as the scale parameter, to use Fisher's terms. These correspond to epistemological needs better than 'true value' and 'standard error' do. But the latter terms are convenient; we have only to remember that 'true value' is not to be understood in an absolute sense, but in the sense that any law relating measures, if it is to be of any use, must be clearly stated, in probability terms, and that a possible way of progress (apparently the only possible way) is to treat the variation as the resultant of a part that would be exactly predictable, given exact statements of the values of certain parameters, and a random error. The law in its naïve form would deal only with the former part. The parameters in this part may be called the true values of the parameters, and the observed values that they would lead to if the random part was neglected the true values. The actual observed values will differ somewhat. By the principle of inverse probability we shall be able then to proceed from the observations to estimates of the true values of the parameters, which, however, will not be exact determinations, but will have ranges of uncertainty corresponding to the fact that the individual random errors in the observations are not definitely known.

In actual fact there are some cases where the normal law of error appears to represent the outstanding variation as well as we can tell. There are others where, though we find that it is probably incorrect when we study a sufficient number of observations, this number is

large, of the order of 500, and the use of the normal law in such cases as if it was correct would not lead to serious mistakes. There are others where it is glaringly wrong, and the only proper treatment is to obtain a sufficient number of observations to give us some idea of what the corresponding distribution of chance can be. Meanwhile we shall consider an important series of generalized laws of error.

**2.3. The Pearson laws.** If we write the normal law of error in the form

$$P(dx | \lambda, \sigma, H) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(x-\lambda)^2}{2\sigma^2}\right\} \frac{dx}{\sigma}, \quad (1)$$

where we have now made the parameters  $\lambda$  and  $\sigma$  explicit (they were formerly understood in  $H$ ), we see that it is an instance of the general form

$$P(dx | H) = y \, dx, \quad (2)$$

where  $y \geq 0$  and the integral of  $y$  over all possible values must be 1. In this case we find easily

$$\frac{1}{y} \frac{dy}{dx} = -\frac{x-\lambda}{\sigma^2}. \quad (3)$$

The law, therefore, has the properties that  $dy/dx$  vanishes in the limit when  $y$  tends to 0, and at one intermediate value of  $x$ , namely,  $\lambda$ . If we consider the generalized form

$$\frac{1}{y} \frac{dy}{dx} = -\frac{x-a}{b_0 + b_1 x + b_2 x^2}, \quad (4)$$

the same will usually hold, but we have two more parameters and shall be able to represent laws of a much wider range of form. They will have one point where  $y$  is stationary; if the range of  $x$  is infinite  $y$  and  $dy/dx$  will tend to zero at the end or ends; if the range is limited in one or both directions there will still be cases where this holds. The integral of (4) can in general be written in the form

$$y = A(x-c_1)^{m_1}(c_2-x)^{m_2}, \quad (5)$$

where  $A$  will be fixed by the condition that the integral of  $y$  is 1, and  $c_1$  and  $c_2$  are the zeros of the denominator in (4). There are three main types of solution and a number of transitional and degenerate cases.

1.  $c_1$  and  $c_2$  imaginary. Then they must be conjugate complexes, and for  $y$  to be real  $m_1$  and  $m_2$  must also be conjugate complexes.  $y$  cannot vanish or become infinite for any real value of  $x$ , and the admissible values of  $x$  range from  $-\infty$  to  $+\infty$ , with a maximum of  $y$

at some intermediate value. Forms with one maximum are designated *bell-shaped* by Pearson. We may write these laws in the forms

$$y = A(x - \lambda - i\beta)^{-m+iq}(x - \lambda + i\beta)^{-m-iq} \\ = (2\beta)^{2m-1} \frac{(m-1+iq)!(m-1-iq)!}{2\pi(2m-2)!} \{(x-\lambda)^2 + \beta^2\}^{-m} \times \\ \times \exp\left(-2q \tan^{-1} \frac{x-\lambda}{\beta}\right). \quad (6)$$

These are Pearson's *Type IV*. In general they are asymmetrical or *skew*, but if  $q = 0$  they reduce to the symmetrical form

$$y = (2\beta)^{2m-1} \frac{(m-1)!(m-1)!}{2\pi(2m-2)!} \{(x-\lambda)^2 + \beta^2\}^{-m} \quad (7)$$

$$= \beta^{2m-1} \frac{(m-1)!}{\pi^{1/2}(m-\frac{3}{2})!} \{(x-\lambda)^2 + \beta^2\}^{-m}, \quad (8)$$

which is Pearson's *Type VII*. In both cases  $m$  must be greater than  $\frac{1}{2}$  for convergence. These laws resemble the normal law in having an infinite range of  $x$  in both directions, which is true of no other Pearson type, but  $y$  falls off less rapidly. With the normal law the expectation of any power of  $x$  is finite; with Type VII that of any even power equal to  $2m-1$  or more is infinite ( $m$  need not be integral); with Type IV expectations of odd powers  $\geq 2m-1$  are also infinite. This is a useful property in representing errors of measurement, since it is usually found, when sufficient observations are available, that there are more outlying large residuals than the normal law would suggest. The fact that these laws, like the normal law, give a non-zero chance of an error greater than any finite amount is an apparent drawback, since we might say that however bad the observations are there is some limit to the error; but to harmonize this belief with the observed distributions would require us to go beyond the range of the Pearson types, which do give satisfactory agreement within the ranges where observations exist.

If  $c_1$  and  $c_2$  are real ( $c_2 > c_1$ ) we must distinguish three cases. (4) has singularities at  $c_1$  and  $c_2$  and the solution is applicable only in ranges that do not include a singularity. Hence we must consider separately cases where the admissible values of  $x$  are less than  $c_1$ , between  $c_1$  and  $c_2$ , or greater than  $c_2$ . The difference between the first and third can be removed by merely reversing the direction of measurement.

2. Admissible values of  $x$  between  $c_1$  and  $c_2$ . We can take the law in the form

$$y = \frac{(m_1 + m_2 + 1)!}{m_1! m_2! (c_2 - c_1)^{m_1 + m_2 + 1}} (x - c_1)^{m_1} (c_2 - x)^{m_2}, \quad (9)$$

which will be possible if both  $m_1$  and  $m_2$  are greater than  $-1$ . If both are positive, the curve is bell-shaped. If  $0 > m_1 > -1$ ,  $y$  is infinite at  $c_1$ . If at the same time  $m_2$  is positive,  $dy/dx$  is negative throughout the range and the curve is called *J-shaped*. In this case  $a$  does not lie between  $c_1$  and  $c_2$ , and is not an admissible value of  $x$ . If  $m_1$  and  $m_2$  are both negative,  $y$  is infinite at both limits and  $a$  lies between them. The curve is then called *U-shaped*. These cases cover Pearson's *Type I*. It will be seen that the possibility of *U-shaped* and *J-shaped* curves gives it greater generality than was originally attempted.

There are several special cases:

$m_1 = m_2$ . The law is then symmetrical. This is Pearson's *Type II*.

Further degenerations give

$m_1 = m_2 = 0$ . This makes  $y$  uniform between  $c_1$  and  $c_2$ , and zero outside that range. This is the *rectangular distribution*, not given a number by Pearson.

$m_1 = m_2 = 1$ . This, with a change of scale and origin, gives  $y \propto 1 - x^2$ , the *parabolic distribution*.

$m_1 = 0$ . This is a *J-shaped* curve with  $y$  proportional to  $(c_2 - x)^{m_2}$  for  $x$  between  $c_1$  and  $c_2$ . This is Pearson's *Type IX*. It starts from a finite ordinate at  $c_1$ .

$m_1 = -m_2$ .  $y$  will be proportional to  $\left(\frac{x - c_1}{c_2 - x}\right)^m$  with  $-1 < m < 1$ .

This is Pearson's *Type XII*. The curve is always *J-shaped*.

3. Admissible values of  $x$  all  $\geq c_2$ . We can take the law in the form

$$y = \frac{(-m_1 - 1)!}{m_2!(-m_1 - m_2 - 2)!(c_2 - c_1)^{m_1 + m_2 + 1}} (x - c_1)^{m_1} (c_2 - x)^{m_2}, \quad (10)$$

where for convergence  $m_2 > -1$ ,  $m_1 + m_2 < -1$ . These are the laws of *Type VI*. If  $m_2 > 0$  they are bell-shaped; if  $m_2 < 0$ , *J-shaped*. They are never *U-shaped*. These laws will give the kind of distribution shown by the times of arrival of a train; there is a concentration at values a little greater than  $c_2$ , values less than  $c_2$  do not occur, and there is a long train of large values, which may rarely occur but are serious when they do.

A particular case is

$m_2 = 0$ . This makes  $y$  proportional to  $(x - c_1)^{m_1}$  for values of  $x$  greater than  $c_2$ ; evidently  $m_1 < -1$ . This gives Pearson's *Types VIII and XI*, which are identical. It starts from a finite ordinate at  $c_2$ .

Types IV, I, and VI, to take them in what seems to me to be their natural order, are the only ones that involve the full number of adjustable parameters, four. There are also three transitional cases between them.

4. There will be a transition from Type I to Type VI expressed by making  $c_2$  in I tend to  $+\infty$  or  $c_1$  in VI to  $-\infty$ . In either case the limiting form is

$$y \propto (x-c)^m e^{-\alpha x} \quad (m > -1, \alpha > 0).$$

This is *Type III*. It resembles Type VI in appearance but is more closely concentrated to small departures from  $c$ . A particular case is

$m = 0$ ; this is *Type X*, an exponential law, which can also be regarded as the transition between Types VIII and IX.

5. The transition from Type VI to Type IV is the case of equal roots, the roots of the denominator in (4) being equal, real, and finite. Then we can write (4) in the form

$$\frac{1}{y} \frac{dy}{dx} = -\frac{\alpha}{x-c} + \frac{\beta}{(x-c)^2},$$

whence 
$$y = A(x-c)^{-\alpha} \exp\left(-\frac{\beta}{x-c}\right).$$

This is *Type V*. To give convergence at  $\infty$ ,  $\alpha$  must be  $> 1$ ; for convergence at  $c$ ,  $\beta > 0$  for any  $\alpha > 1$ . It is always bell-shaped, since  $y$  must vanish at  $x = c$ . Otherwise it resembles Type VI. It differs from Type III in the interchange of the two types of convergence at the extremes; indeed, the change of  $(x-c)$  to  $(x-c)^{-1}$  transforms one into the other.

6. The transition from Type IV to Type I requires the roots to be  $\pm\infty$ ; then  $b_1$  and  $b_2$  both vanish and we are back to the normal law.

This analysis covers the range of the Pearson types, and is, I think, considerably shorter and more systematic than has been given previously. My own experience with them has been rather small, though I have had to deal with Types II, III, VII, and VIII. For purposes of exposition I think it would be a great convenience if those who use them extensively could agree on a more systematic numbering in place of the present haphazard one, which places III, the transition between I and VI, between II, which is the symmetrical case of I, and IV, which is a different main type from any; and VI, a main type, between V, a transitional case, and VII, a degenerate case of IV. I should suggest the following.



Main types	Pearson's number	Special cases	Number	
			Pearson's	Suggested
1	IV	$q = 0$	VII	1a
2	I	$m_1 = m_2$	II	2a
		$m_1 = m_2 = 0$	Rect.	2b
		$m_1 = m_2 = 1$	Parab.	2c
		$m_1 = 0$	IX	2d
		$m_1 = -m_2$	XII	2e
3	VI	$m_2 = 0$	VIII	3a
Transitions				
2 to 3	III	$m = 0$	X	23a
3 to 1	V			
1 to 2	Normal			

This covers the whole range with the exception of XI, which is a mere rewriting of VIII. I think that special numbers for the rectangular and parabolic laws are worth while as they are likely to be at least as important as XII in practice, and the rectangular law has great theoretical interest. Both, like the normal law, involve only a scale parameter and a location parameter. The main types involve two others. The rest involve three parameters in all.

It may be remarked that Pearson distinguished Types I and VI according as the roots are real and of opposite sign or real and of like sign. This appears to make the type depend on the arbitrary position of the origin. The important point is whether the admissible values of  $x$  lie between the roots or not. In fact Pearson does make his decision according to the latter criterion.

**2.4. The negative binomial law.** Suppose that a distribution of chance follows the Poisson law

$$P(l | rH) = \frac{r^l}{l!} e^{-r} \quad (1)$$

but that  $r$  itself is unknown, having a distribution of chance given by the Type III law

$$P(dr | H) = \beta^{\alpha+1} \frac{r^\alpha}{\alpha!} e^{-\beta r} dr \quad (2)$$

(where, since  $\alpha$  may be fractional, we must understand  $\alpha!$  to be defined by  $\alpha! = \int_0^\infty t^\alpha e^{-t} dt$ ). Then

$$P(l, dr | H) = \beta^{\alpha+1} \frac{r^{l+\alpha}}{l! \alpha!} e^{-(1+\beta)r} dr. \quad (3)$$

To get the total probability for any value of  $l$ , we must add for all

possible values of  $r$ ; which means in this case that we must integrate. Then

$$P(l|H) = \int_0^{\infty} \frac{\beta^{\alpha+1} r^{l+\alpha}}{l! \alpha!} e^{-(1+\beta)r} dr = \frac{\beta^{\alpha+1} (l+\alpha)!}{(1+\beta)^{l+\alpha+1} l! \alpha!}. \quad (4)$$

Apart from the factor  $\left(\frac{\beta}{\beta+1}\right)^{\alpha+1}$ , this is the coefficient of  $x^l$  in the expansion of  $\left(1 - \frac{x}{\beta+1}\right)^{-\alpha-1}$ . The sum over all values of  $l$  is 1, as it must be since the conditions stated are exhaustive. If we put

$$\frac{\beta}{\beta+1} = 1-a,$$

we have 
$$P(l|H) = (1-a)^{\alpha+1} \frac{(\alpha+l)!}{\alpha! l!} a^l, \quad (5)$$

which puts the negative binomial form more clearly in evidence. This result is due to M. Greenwood and G. U. Yule.† The immediate application was to problems of factory accidents. The conditions of the Poisson law were satisfied in respect of the total chance of an accident in a factory in a given period being the sum of a large number of small chances, but it was not clear that these chances were the same for all employees. The chance of a particular workman having an accident on a particular day, for instance, would have to be regarded as the analogue of  $x$  in the derivation of the Poisson law, and the number of days in the period considered as the analogue of  $n$ . Then for each individual the chances of 0, 1, 2, ... accidents in the period would follow a Poisson law—subject to the condition that having one accident does not stimulate him to have another—and if the values of  $r = nx$  for the different workmen are distributed, as nearly as can be for a finite number, in a Type III law, the negative binomial follows as the resultant for all workmen.

The following alternative development shows that the condition that the probabilities of accidents to the same workman must be independent is not strictly necessary. It can at any rate be replaced by other conditions. Suppose that the total number of events is recorded, but that in fact some of the events are composite, two or more being associated. These are each only one independent event, but will be counted as two or more each in the totals. Let  $r_1, r_2, \dots$  be the appropriate values of  $r$  for the simple, double, ... events in the interval considered. Each type

† *J. R. Stat. Soc.* 83, 1920, 255–79.

separately will satisfy the Poisson rule, and the chance that there will be  $m_1$  simple,  $m_2$  double events, and so on, will be

$$P(m_1, m_2, \dots | r_1, r_2, \dots, H) = \frac{r_1^{m_1} r_2^{m_2}}{m_1! m_2!} \dots \exp\{-(r_1 + r_2 + \dots)\}. \quad (6)$$

The probability that the total number of events as counted will be  $m$  is the sum of these expressions, subject to

$$m_1 + 2m_2 + 3m_3 + \dots = m. \quad (7)$$

But this sum is the coefficient of  $x^m$  in the expansion of

$$f(x) = \exp(r_1 x + r_2 x^2 + \dots - r_1 - r_2 - \dots). \quad (8)$$

Now in practice, if we have no record of the individual events, there will not be much hope of determining the  $r$ 's separately. But if we want to find a law that will take into account the extra complication we must have at least one new parameter, though there may not be much point in introducing more than one. Let us take the form:

$$r_s = r_1 a^{s-1}/s. \quad (9)$$

Then

$$\begin{aligned} \log f(x) &= r_1 x(1 + \tfrac{1}{2}ax + \tfrac{1}{3}a^2x^2 + \dots) - r_1(1 + \tfrac{1}{2}a + \dots) \\ &= (r_1/a)\{-\log(1-ax) + \log(1-a)\}, \end{aligned} \quad (10)$$

$$f(x) = \left(\frac{1-a}{1-ax}\right)^{r_1/a}, \quad (11)$$

and the coefficient of  $x^m$  is

$$P(m | r_1, a, H) = (1-a)^{r_1/a} \frac{r_1}{a} \left(\frac{r_1}{a} + 1\right) \dots \left(\frac{r_1}{a} + m - 1\right) \frac{a^m}{m!}, \quad (12)$$

which again is a negative binomial law, with  $r_1/a$  replacing the  $\alpha + 1$  of Greenwood and Yule's derivation.†

It is convenient to take the law in the form

$$P(m | r, n, H) = \left(\frac{n}{n+r}\right)^n \frac{n(n+1)\dots(n+m-1)}{m!} \left(\frac{r}{n+r}\right)^m. \quad (13)$$

When  $n \rightarrow \infty$  this tends to the Poisson law with parameter  $r$ . We shall see later that it has other advantages. The series converges for all positive  $n$ . The expectations of  $m$  and  $m(m-1)$  are  $r$  and  $(1+1/n)r^2$ . That of  $(m-r)^2$  is  $r+r^2/n$ . When  $n \rightarrow 0$ , all the chances of non-zero  $m$  tend to 0, while that of  $m$  being zero tends to 1. In the latter case as we approach the limit, keeping  $r$  fixed, the chances of  $m$  become more and more widely spread to wide values, and the concentration at 0 is needed to keep the total expectation equal to  $r$ . Thus the negative

† This derivation has already been given by R. Lüders, *Biometrika*, 26, 1934, 108-28.

binomial law, for small  $n$ , will resemble the distribution of the scores of a first-class cricket or billiards player, whose commonest score may be 0 though his average is about 60. On the Poisson law the commonest score and the average should approximately agree, and the chance of a score of 1 would be 60 times that of a score 0.

Here we have a case where two different types of departure from the Poisson law both lead to results of the same form, and modify it in the same direction. If the law is nevertheless found to agree with the facts, it is reasonable to reject both types of departure. Thus the agreement of the data about deaths from kicks of a horse in the Prussian army may be taken to mean both (1) that nobody can be killed twice by the kick of a horse, (2) that the fact that one man has been so killed does not indicate an extra liability for others in the same unit to be. The agreement in the radioactivity data would mean that (1) the chances of disintegration of different atoms of the same radioactive substance are approximately equal, (2) the disintegration of one atom does not lead immediately to the disintegration of another.

**2.5. Correlation.** This can be treated on lines analogous to the derivation of the normal law from the binomial. Suppose that two quantities  $x$  and  $y$  are to be measured simultaneously, and that there are  $m+n$  independent component variations, each contributing  $\pm\alpha$  to  $x$  and  $\pm\beta$  to  $y$ .  $m$  of them are constrained to give the same sign in both  $x$  and  $y$ ,  $n$  to give opposite signs. Suppose that in a particular case the number making positive contributions to  $x$  that give the same sign is  $p$ , the number giving opposite signs  $q$ . Then

$$x = p\alpha - (m-p)\alpha + q\alpha - (n-q)\alpha = (2p-m)\alpha + (2q-n)\alpha, \quad (1)$$

$$y = p\beta - (m-p)\beta - q\beta + (n-q)\beta = (2p-m)\beta - (2q-n)\beta. \quad (2)$$

We are taking each component to be as likely as not to give a positive contribution to  $x$ . Then

$$\begin{aligned} P(p, q \mid m, n, \alpha, \beta, H) &= 2^{-m-n} {}^mC_p {}^nC_q \\ &= \frac{2}{\pi\sqrt{mn}} \exp\left\{-\frac{2}{m}(p-\tfrac{1}{2}m)^2 - \frac{2}{n}(q-\tfrac{1}{2}n)^2\right\} \end{aligned} \quad (3)$$

by the previous argument. We have to transform to the observed variables  $x$  and  $y$ . Now

$$\frac{\partial(x, y)}{\partial(p, q)} = 8\alpha\beta; \quad 2p-m = \frac{1}{2}\left(\frac{x}{\alpha} + \frac{y}{\beta}\right); \quad 2q-n = \frac{1}{2}\left(\frac{x}{\alpha} - \frac{y}{\beta}\right). \quad (4)$$

Remembering that  $p$  and  $q$  are capable of integral values only, and that

the total chance in any region must be the same whether the observation is expressed in terms of  $p$  and  $q$  or of  $x$  and  $y$ , we see that we must replace the sum with regard to  $p$  and  $q$  by the integral with regard to  $dx dy / 8\alpha\beta$ . Hence

$$P(dx dy | m, n, \alpha, \beta, H) = \frac{dx dy}{4\pi\alpha\beta\sqrt{mn}} \exp\left\{-\frac{1}{8m}\left(\frac{x}{\alpha} + \frac{y}{\beta}\right)^2 - \frac{1}{8n}\left(\frac{x}{\alpha} - \frac{y}{\beta}\right)^2\right\}. \quad (5)$$

Now put

$$(m+n)\alpha^2 = \sigma^2; \quad (m+n)\beta^2 = \tau^2; \quad (m-n)\alpha\beta = \rho\sigma\tau. \quad (6)$$

Then we find

$$P(dx dy | m, n, \alpha, \beta, H) = \frac{dx dy}{2\pi\sigma\tau\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{x^2}{\sigma^2} - \frac{2\rho xy}{\sigma\tau} + \frac{y^2}{\tau^2}\right)\right\}, \quad (7)$$

so that the four original parameters are now reduced to three, and we can assert that this is also equal to  $P(dx dy | \sigma, \tau, \rho, H)$ . Of course, everything that can be said against the normal law of error for one variable can be said twice against this form, which is the generalization to two variables. But on the other hand the chief thing that can be said in favour of the normal law, that of all laws that are anywhere near the truth it is far the easiest to apply, can also be said with greater force of normal correlation. The new parameter  $\rho$  is called the *correlation coefficient*.

The law (7) was obtained first by Sir Francis Galton empirically, by studying observed frequencies.† As Pearson remarks:‡ ‘That Galton should have evolved all this from his observations is to my mind one of the most noteworthy scientific discoveries arising from pure analysis of observations.’ Galton had not, at this stage, noticed that negative correlations exist, since he remarks: ‘Two variable organs are said to be correlated when the variation of one is accompanied on the average by more or less variation of the other, and in the same direction,’§ and he speaks of correlation arising when two variations are the resultant of several causes, some common to both and some independent. The above analysis permits negative correlations. The more restricted one, however, is often valid and leads in particular to an account of intra-class correlation.

By integration we find

$$P(dx | \sigma, \tau, \rho, H) = \frac{1}{\sqrt{(2\pi)\sigma}} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx. \quad (8)$$

† *B.A. Report*, Aberdeen, 1885.

‡ *Biometrika*, **13**, 1920, 25–45. This is a most interesting historical study.

§ *Proc. Roy. Soc.* **45**, 1889, 135.

Therefore

$$P(dy | \sigma, \tau, \rho, x, H) = \frac{P(dxdy | \sigma, \tau, \rho, H)}{P(dx | \sigma, \tau, \rho, H)} \\ = \frac{1}{\sqrt{(2\pi)\tau}\sqrt{(1-\rho^2)}} \exp\left\{-\frac{1}{2\tau^2(1-\rho^2)}\left(y - \frac{\rho\tau x}{\sigma}\right)^2\right\} dx. \quad (9)$$

That is, the probability of  $x$  is normally distributed with standard error  $\sigma$ , and for given  $x$  the probability of  $y$  is normally distributed about  $\rho\tau x/\sigma$  with standard error  $\tau\sqrt{(1-\rho^2)}$ . The line  $y = \rho\tau x/\sigma$  is known as the *line of regression* of  $y$  on  $x$ . Similarly the probability of  $y$  is normally distributed with standard error  $\tau$ , and that of  $x$  given  $y$  is normally distributed about  $x = \rho\sigma y/\tau$ , the line of regression of  $x$  on  $y$ . The lines of regression coincide only if  $\rho = \pm 1$ .

The expectations of  $x^2$ ,  $y^2$ , and  $xy$ , given  $\sigma$ ,  $\rho$ ,  $\tau$ ,  $H$ , are respectively  $\sigma^2$ ,  $\tau^2$ ,  $\rho\sigma\tau$ .

**2.6. The characteristic function.** Suppose that on a given law the chance of the variable  $x$  being less than an assigned value is  $f(x)$ . Then the expectation of any function  $\lambda(x)$  of  $x$  is  $\int \lambda(x) df(x)$  over the range of  $x$ ; in which we must understand a Stieltjes integral if  $f(x)$  has discontinuities. These, if any, will all be positive jumps. The characteristic function  $\Omega(\kappa)$  is defined as the expectation of  $e^{\kappa x}$ , where  $\kappa$  is purely imaginary; thus

$$\Omega(\kappa) = \int e^{\kappa x} df(x) \quad (1)$$

and  $|\Omega(\kappa)| \leq 1$ . The integral is absolutely convergent because  $\int df(x)$  converges.

The integral

$$\frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} \frac{e^{\kappa x}}{\kappa} (e^{-\kappa x_1} - e^{-\kappa x_2}) d\kappa \quad (x_1 < x_2),$$

in which the path is a line parallel to the imaginary axis on the positive side, is equal to 1 if  $x_1 < x < x_2$ , and zero if  $x < x_1$  or  $x > x_2$ , being the difference of two Heaviside unit functions. If we replace the path by the imaginary axis, except for a small semicircle about the origin, the integral is unaltered. Also the integral about the small semicircle tends to zero in the limit and the integrand is continuous. Hence we may replace the path by the imaginary axis, and

$$\frac{1}{2\pi i} \int_{-i\infty}^{i\infty} \frac{e^{\kappa x}}{\kappa} (e^{-\kappa x_1} - e^{-\kappa x_2}) d\kappa = \begin{cases} 1 & (x_1 < x < x_2), \\ 0 & (x < x_1, x_2 < x). \end{cases} \quad (2)$$

Now consider the sum

$$\sum \frac{1}{2\pi i} \int_{-i\infty}^{i\infty} \frac{e^{\kappa \xi'_r}}{\kappa} (e^{-\kappa x_1} - e^{-\kappa x_2}) \{f(\xi_{r+1}) - f(\xi_r)\} d\kappa \quad (3)$$

over  $r$ , the ranges from  $\xi_r$  to  $\xi_{r+1}$  being so chosen that all points of discontinuity of  $f(\xi)$  lie within them, and  $\xi'_r$  being some value between  $\xi_r$  and  $\xi_{r+1}$ . On integrating with regard to  $\kappa$ , terms for  $\xi'_r$  not between  $x_1$  and  $x_2$  vanish, while those between them contribute

$$\sum_{x_1}^{x_2} \{f(\xi_{r+1}) - f(\xi_r)\} \rightarrow f(x_2) - f(x_1) \quad (4)$$

in the limit when the intervals become indefinitely short. But the limit of the sum is by definition the Stieltjes integral

$$\frac{1}{2\pi i} \int_{x=-\infty}^{\infty} df(x) \int_{-i\infty}^{i\infty} \frac{1}{\kappa} (e^{-\kappa x_1} - e^{-\kappa x_2}) e^{\kappa x} d\kappa = \frac{1}{2\pi i} \int_{-i\infty}^{i\infty} \frac{1}{\kappa} (e^{-\kappa x_1} - e^{-\kappa x_2}) \Omega(\kappa) d\kappa, \quad (5)$$

by inverting the order of integration, which is easily shown to be valid.

When  $f(x)$  is differentiable this leads to a case of Fourier's integral theorem

$$\frac{df(x)}{dx} = \frac{1}{2\pi i} \int_{-i\infty}^{i\infty} e^{-\kappa x} \Omega(\kappa) d\kappa. \quad (6)$$

Similarly, if  $\epsilon_1, \epsilon_2, \dots, \epsilon_k$  are a set of variables whose chances are independent and follow laws given by  $f_1(\epsilon_1), f_2(\epsilon_2), \dots$  it can be shown that the chance that

$$x_1 < \sum \epsilon_r < x_2$$

is

$$\begin{aligned} & \frac{1}{2\pi i} \int_{-i\infty}^{i\infty} d\kappa \left[ \int_{\epsilon_r=-\infty}^{\infty} \right]^k \frac{1}{\kappa} e^{\kappa \sum \epsilon_r} (e^{-\kappa x_1} - e^{-\kappa x_2}) df_1 \dots df_k, \\ &= \frac{1}{2\pi i} \int_{-i\infty}^{i\infty} \frac{1}{\kappa} (e^{-\kappa x_1} - e^{-\kappa x_2}) \Omega_1(\kappa) \Omega_2(\kappa) \dots \Omega_k(\kappa) d\kappa, \end{aligned} \quad (7)$$

where the  $\Omega$ 's are the characteristic functions corresponding to the  $f$ 's. Hence the characteristic function of the sum of a set of variables following independent laws of chance is the product of their separate characteristic functions.

The characteristic function is intimately related to the expectations of the powers of  $x$ , where these exist. If we write

$$\mu_m = \int x^m df(x), \quad (8)$$

we can call  $\mu_m$  the  $m$ th moment of the law about the origin. If moments up to order  $m$  exist, we can differentiate (1)  $m$  times under the integral sign with regard to  $\kappa$ , and then for  $\kappa = 0$

$$\frac{d^m}{d\kappa^m} \Omega(\kappa) = \mu_m. \quad (9)$$

Thus, by Taylor's theorem,

$$\Omega(\kappa) = 1 + \mu_1 \kappa + \mu_2 \frac{\kappa^2}{2!} + \dots + \mu_m \frac{\kappa^m}{m!} + o(\kappa^m), \quad (10)$$

even though the complete Taylor series may not exist. For this reason  $\Omega(\kappa)$  is also called the moment-generating function. If we take the origin of  $x$  at its expectation,  $\mu_1$  will be 0. Inspection of (1) shows that decreasing all values of  $x$  by  $\mu_1$  will multiply  $\Omega(\kappa)$  by  $e^{-\kappa\mu_1}$ , and therefore if  $\Omega_0(\kappa)$  is the characteristic function of  $x - \mu_1$ ,

$$\Omega_0(\kappa) = e^{-\kappa\mu_1} \Omega(\kappa).$$

The coefficients of  $\kappa^n/n!$  in the expansion of  $\log \Omega(\kappa)$  are called the semi-invariants or cumulants, when they exist, since the second and higher ones are independent of the origin and are additive for the sum of several variables. Also if  $y$  has a probability law  $g(y)$  such that  $g(y) = f(x)$  if  $y = ax$ , where  $a$  is constant, the characteristic function of  $g(y)$  is

$$E(\kappa) = \int e^{\kappa y} dg(y) = \int e^{a\kappa x} df(x) = \Omega(a\kappa). \quad (11)$$

The moment and the semi-invariant of  $g(y)$  of order  $m$  are  $a^m$  times those of  $f(x)$ .

If  $\Omega(\kappa)$  can be expanded in powers of  $\kappa$ , it will follow that the series represents an analytic function near  $\kappa = 0$ . But if any moment of the law diverges, the integral (1) defining  $\Omega(\kappa)$  will not exist for  $\kappa$  on at least one side of the imaginary axis, however close to it, since the integral will contain a factor  $e^{cx}$ , where  $c$  is real and not zero. Thus the integral will define a function *only* for purely imaginary values of  $\kappa$ . It may be the value on the imaginary axis of some function analytic in the half-plane, but such a function, if it exists, will not be given off the axis by the integral. This applies to laws of Pearson's Types IV, VII, and VI.

The integral may exist for all real  $\kappa$ ; this applies in all cases where the law has a finite range, such as the binomial and Type I laws. It is also true for the normal law. In that case the integral will exist for all  $\kappa$  and be uniformly convergent in any bounded region of the  $\kappa$  plane. It can therefore be integrated under the integral sign about any contour



in the  $\kappa$  plane, and this integral will be 0 since  $\int_C e^{\kappa x} d\kappa = 0$ . Hence by Morera's theorem†  $\Omega(\kappa)$  is an analytic function within any contour in the  $\kappa$  plane, and must therefore be an integral function.‡ Then  $\Omega(\kappa)$  is expandible in powers of  $\kappa$  over the entire plane.

There are cases where the integral exists for some complex values of  $\kappa$  and not for others; for instance, the median law

$$df = \frac{1}{2} \exp(-|x|/a) dx/a.$$

Within the belt  $-1/a < R(\kappa) < 1/a$  the integral will define an analytic function. Outside this belt it diverges.

Thus we have two main types of case. If all the moments of the law exist and the expectations of  $e^{\pm cx}$  also exist, where  $c$  is some real quantity,  $\Omega(\kappa)$  will be analytic near 0 and the coefficient of  $\kappa^n$  will be  $\mu_n/n!$  for all  $n$ . If moments up to order  $m$  converge, but those of higher orders diverge, the integral does not define a function except for purely imaginary values of  $\kappa$ . Its derivatives at  $\kappa = 0$  for imaginary  $\kappa$  will give the moments correctly up to order  $m$ ; but higher derivatives, if they exist, will not give the higher moments. We shall see that they do not necessarily exist.

**2.61.** The characteristic function is sometimes useful for actually calculating the moments. Thus consider the binomial law, according to which the chance of a sampling number less than  $l$  is

$$f(l) = \sum_0^{l-1} {}^nC_l x^l (1-x)^{n-l}. \quad (1)$$

$$\text{Then} \quad \Omega(\kappa) = \sum_{l=0}^n {}^nC_l x^l (1-x)^{n-l} e^{\kappa l} = (xe^{\kappa} + 1 - x)^n. \quad (2)$$

The coefficient of  $\kappa$  is  $nx$ , which is therefore the expectation of  $l$ . The moments about  $nx$  can then be derived by considering

$$\begin{aligned} \Omega_0(\kappa) &= (1-x + xe^{\kappa})^n e^{-n\kappa x} \\ &= \exp \left\{ \frac{1}{2!} n\kappa^2 xy + \frac{n\kappa^3}{3!} xy(y-x) + \frac{n\kappa^4}{4!} xy(1-6xy) + \dots \right\} \\ &= 1 + \frac{n\kappa^2}{2!} xy + \frac{n\kappa^3}{3!} xy(y-x) + \frac{\kappa^4}{4!} \{3n^2 x^2 y^2 + nxy(1-6xy)\} + \dots, \quad (3) \end{aligned}$$

where  $y = 1-x$ ; whence the moments to order 4 about the mean are

$$\mu_2 = nxy; \quad \mu_3 = nxy(y-x); \quad \mu_4 = 3n^2 x^2 y^2 + nxy(1-6xy). \quad (4)$$

† E. C. Titchmarsh, *Theory of Functions*, 1932, p. 82.

‡ I am indebted to Professor Littlewood for calling my attention to this point, in answer to a query.

Pearson's parameters  $\sqrt{\beta_1}$  and  $\beta_2$  are given by

$$\sqrt{\beta_1} = \frac{\mu_3}{\mu_2^{3/2}} = \frac{y-x}{(nxy)^{1/2}}; \quad \beta_2 = \frac{\mu_4}{\mu_2^2} = 3 + \frac{1-6xy}{nxy}. \quad (5)$$

$\sqrt{\beta_1}$  and  $\beta_2$  are the characteristic form parameters (as distinct from those of location and scale) used by him in fitting his characteristic laws and other types of law. They are otherwise useful as a general indication of the features of a law. If  $x < \frac{1}{2}$ , the positive sign of  $\sqrt{\beta_1}$  indicates the skewness due to the longer range on the upper side of the mean. If  $x = \frac{1}{2}$ , the law is symmetrical and  $\beta_2 = 3 - 2/n$ . In the limit when  $n$  is large and the law tends to the normal, therefore, the fourth moment tends to three times the square of the second. The fact that  $\beta_2 < 3$  for the symmetrical binomial is an indication of the effect of the finite range. The law is lower in the middle and at the tails than the normal law with the same  $\mu_2$ .

In (2) put  $x = r/n$  and let  $n$  tend to infinity; then the law tends to the Poisson form. In this case the mean of the law is  $r$ ; shifting the origin to the mean we have

$$\Omega_0(\kappa) = \exp\{r(e^\kappa - 1 - \kappa)\} = \exp r \left( \frac{\kappa^2}{2!} + \frac{\kappa^3}{3!} + \dots \right) \quad (6)$$

$$= 1 + \frac{r\kappa^2}{2!} + \frac{r\kappa^3}{3!} + (3r^2 + r) \frac{\kappa^4}{4!} + \dots, \quad (7)$$

$$\text{whence} \quad \mu_2 = r, \quad \mu_3 = r, \quad \mu_4 = 3r^2 + r. \quad (8)$$

The semi-invariants are all equal to  $r$ , by (6).

For the negative binomial law

$$\Omega(\kappa) = \left( \frac{n}{n+r} \right)^n \sum \frac{n(n+1)\dots(n+m-1)}{m!} \left( \frac{r}{n+r} \right)^m e^{m\kappa} \quad (9)$$

$$\begin{aligned} &= \left( \frac{n}{n+r} \right)^n \left( 1 - \frac{r}{n+r} e^\kappa \right)^{-n} \\ &= \left\{ 1 - \frac{r(e^\kappa - 1)}{n} \right\}^{-n}. \end{aligned} \quad (10)$$

The coefficient of  $\kappa$  in the expansion is  $r$ , which is the expectation of  $m$ ; and

$$\Omega_0(\kappa) = \left\{ 1 - \frac{r(e^\kappa - 1)}{n} \right\}^{-n} e^{-r\kappa}, \quad (11)$$

$$\log \Omega_0(\kappa) = \kappa^2 \left( \frac{r}{2} + \frac{r^2}{2n} \right) + \kappa^3 \left( \frac{r}{6} + \frac{r^2}{2n} + \frac{r^3}{3n^2} \right) + \kappa^4 \left( \frac{r}{24} + \frac{7}{24} \frac{r^2}{n} + \frac{r^3}{2n^2} + \frac{r^4}{4n^3} \right) + \dots \quad (12)$$

The second moment is therefore  $r+r^2/n$ , as we found directly in 2.4; the third and fourth are

$$\mu_3 = r + \frac{3r^2}{n} + \frac{2r^3}{n^2}; \quad \mu_4 = r + r^2\left(3 + \frac{7}{n}\right) + r^3\left(\frac{6}{n} + \frac{12}{n^2}\right) + r^4\left(\frac{3}{n^2} + \frac{6}{n^3}\right).$$

For the normal law

$$f(x) = \frac{1}{\sqrt{(2\pi)\sigma}} \int_{-\infty}^x \exp\left(-\frac{x^2}{2\sigma^2}\right) dx$$

$$\text{we find easily} \quad \Omega(\kappa) = \exp\left(\frac{1}{2}\sigma^2\kappa^2\right). \quad (13)$$

All the moments converge, and

$$\mu_{2m} = \frac{(2m)!}{2^m m!} \sigma^{2m}; \quad \mu_{2m+1} = 0. \quad (14)$$

For the median law

$$df = \frac{1}{2} \exp\left(-\frac{|x|}{a}\right) \frac{dx}{a} \quad (15)$$

$$\text{we find} \quad \Omega(\kappa) = \frac{1}{1-a^2\kappa^2}. \quad (16)$$

The second moment is  $2a^2$ , as we can see at once otherwise.

For the binomial, Poisson, and normal laws all the moments exist and the characteristic function is an integral function. For the negative binomial and the median law all the moments exist, but the characteristic function has poles and is not defined over the whole  $\kappa$  plane by 2.6(1).

**2.62.** Consider now a case where the second moment is infinite, the Cauchy distribution (the Type VII law with index 1)

$$\frac{df}{dx} = \frac{1}{\pi(1+x^2)}. \quad (1)$$

The integral for  $\Omega(\kappa)$  must be found by contour integration. When  $I(\kappa)$  is positive the infinite semicircle must be taken on the positive side of the axis of  $x$ , and the contour encloses the pole at  $x = i$ . When  $I(\kappa)$  is negative, on the other hand, the suitable contour encloses the pole at  $-i$ . Thus  $\Omega(\kappa)$  has different analytic forms according to the sign of  $I(\kappa)$ . They are

$$\Omega(\kappa) = \begin{cases} e^{i\kappa} & \{I(\kappa) > 0\}, \\ e^{-i\kappa} & \{I(\kappa) < 0\}. \end{cases} \quad (2) \quad (3)$$

The first derivative of  $\Omega(\kappa)$  does not exist at  $\kappa = 0$ , and no function analytic in any region about  $\kappa = 0$  can represent  $\Omega(\kappa)$ .

For the Type VII law with index 2,

$$\frac{df}{dx} = \frac{2}{\pi(1+x^2)^2}, \quad (4)$$

we find similarly

$$\Omega(\kappa) = \begin{cases} (1-i\kappa)e^{i\kappa} & \{I(\kappa) > 0\}, \\ (1+i\kappa)e^{-i\kappa} & \{I(\kappa) < 0\}. \end{cases} \quad (5)$$

Derivatives to order 2 are continuous at  $\kappa = 0$ , corresponding to the existence of the second moment. But the third derivative at  $\kappa = 0$  has different values on the two sides, and  $\Omega(\kappa)$  is not the form taken by any function analytic in a region about  $\kappa = 0$ .

**2.63. The central limit theorem.** The interest of 2.6 (7) lies in its relation to the resultant of a number of independent disturbances. In many cases, if the number is large, it can be shown that the chance of the resultant is approximately normally distributed. We may notice, first, that if there are two components both following the normal law with standard errors  $\sigma$  and  $\tau$ , the respective values of  $\Omega(\kappa)$  will be  $e^{1/2\kappa^2\sigma^2}$  and  $e^{1/2\kappa^2\tau^2}$ ; and by 2.6 (7) the characteristic function of their sum is  $\exp \frac{1}{2}\kappa^2(\sigma^2 + \tau^2)$ . Hence the distribution of the chance for the sum is normal with standard error  $(\sigma^2 + \tau^2)^{1/2}$ . This can be extended to the composition of any number of normal errors. This principle is called the reproductive property of the normal law.

If for each component  $\epsilon_r$  we take the origin at the expectation of  $\epsilon_r$ , and all the second moments about this origin are 1, we have by 2.6 (10)

$$\Omega_r(\kappa) = 1 + \frac{1}{2}\kappa^2 + o(\kappa^2). \quad (1)$$

If instead we consider  $\epsilon_r/\sqrt{k}$ , the second moment is divided by  $k$ , and by 2.61 (11)

$$\Omega_r(\kappa) = 1 + \frac{\kappa^2}{2k} + o\left(\frac{\kappa^2}{2k}\right) \quad (2)$$

It is to be noticed that  $\Omega_r$  is a function of  $\kappa/\sqrt{k}$  and therefore the remainder term, for any  $\kappa$ , is small compared with  $1/k$  for  $k$  large. Then the characteristic function of  $\sum_{r=1}^k \epsilon_r/\sqrt{k}$  will be

$$\Omega(\kappa) = \Omega_1(\kappa)\Omega_2(\kappa)\dots\Omega_k(\kappa) = \left\{1 + \frac{\kappa^2}{2k} + o\left(\frac{\kappa^2}{2k}\right)\right\}^k \quad (3)$$

if all components follow the same law. But even if they do not we shall have

$$\begin{aligned} \log \Omega(\kappa) &= \sum \log \Omega_r(\kappa) \\ &= \sum_{r=1}^k \log \left\{1 + \frac{\kappa^2}{2k} + o\left(\frac{\kappa^2}{2k}\right)\right\}, \end{aligned} \quad (4)$$

and the differences between the laws appear only in the terms  $o(\kappa^2/2k)$ . If then  $k \rightarrow \infty$ ,  $\log \Omega(\kappa) \rightarrow \frac{1}{2}\kappa^2$ , and in the limit

$$\Omega(\kappa) = \exp(\frac{1}{2}\kappa^2). \quad (5)$$

The chance that  $\sum \epsilon_r/\sqrt{k}$  will be between  $x_1$  and  $x_2$  is therefore

$$\frac{1}{2\pi i} \int_{-i\infty}^{i\infty} \frac{1}{\kappa} (e^{-\kappa x_1} - e^{-\kappa x_2}) e^{\frac{1}{2}\kappa^2} d\kappa. \quad (6)$$

This is differentiable, and the derivative gives the probability density

$$\frac{1}{2\pi i} \int_{-i\infty}^{i\infty} \exp(\frac{1}{2}\kappa^2 - \kappa x) d\kappa = \frac{1}{\sqrt{(2\pi)}} \exp(-\frac{1}{2}x^2). \quad (7)$$

Thus the probability distribution of the sum of  $k$  component variations, all following independent laws of chance with second moments  $1/k$ , will tend in the limit as  $k$  becomes large to the normal law with standard error 1.

It is not quite obvious that if a sequence of characteristic functions tends to a limit, that limit is the characteristic function of the limit, if any, of the corresponding laws. It is proved by H. Cramér that the convergence is uniform and therefore that the passage from (5) to (6) is justified.

If the components have not all the same second moment, the result is not necessarily true; Whittaker and Robinson† give a striking example to the contrary.

The above argument is given, with more attention to mathematical detail, by H. Cramér.‡ The important point is that it does not assume the existence of moments above the second. The derivation of the normal law on similar principles, given by Whittaker and Robinson and reproduced with minor changes in my *Scientific Inference*, is no longer of much interest. For it was assumed in the course of the proof that the functions  $\Omega_r(\kappa)$  are all expansible in powers of  $\kappa$ , with coefficients given by the moments, which can be true only if all the moments are finite. The resultant of several components, each satisfying the normal law, itself satisfies the law exactly. The extreme departure from the normal law for each component that would make all the moments finite is one where the chance is concentrated in two values, since any further spread would amount to a smoothing of the distribution and make it more like the normal. But we already know that the resultant

† *Calculus of Observations*, p. 178.

‡ *Random Variables and Probability Distributions*, 1937.

of several components in this case would give a binomial law and would be approximately normal if there were several components. The proof, therefore, added little to what was already obvious.

The argument has been extended by various writers to the case where the components  $\epsilon_r$  follow independent laws  $f_r(\epsilon_r)$  with second moments  $\mu_{2,r}$  about 0, provided that as  $n \rightarrow \infty$ ,  $M_n = \sum_{r=1}^n \mu_{2,r} \rightarrow \infty$  and for all  $r$ ,  $\mu_{2,r}/M_n \rightarrow 0$ . In other words, the second moment for the sum tends to infinity but the largest proportional contribution from a component tends to zero. Then if  $x_n = \sum_1^n \epsilon_r$ ,

$$P\left(\frac{x_n}{\sqrt{M_n}} < \xi \mid H\right) \rightarrow \int_{-\infty}^{\xi} \frac{1}{\sqrt{(2\pi)}} e^{-1/2 u^2} du.$$

Details are given by Kendall.†

**2.64.** If one or more of the components have an infinite  $m$ th moment, ( $m > 2$ ) and the number of components is finite, the normal law can be approximate only in a rather peculiar sense, for it makes all the moments finite, whereas in such a case the  $m$ th moment for the resultant is infinite. An investigation of a special case is desirable to see what this sense can be. But it is convenient to take first the Cauchy law of 2.62. For the resultant of  $k$  components

$$\Omega(\kappa) = \begin{cases} e^{k i \kappa} & \{I(\kappa) > 0\}, \\ e^{-k i \kappa} & \{I(\kappa) < 0\}, \end{cases} \quad (1)$$

(2)

and the probability density is

$$\frac{1}{2\pi i} \int_{-i\infty}^{i\infty} e^{-\kappa x} \Omega(\kappa) d\kappa = \frac{k}{\pi(k^2 + x^2)}, \quad (3)$$

which is of the form for one component, but with the scale multiplied by  $k$ . The mean of  $k$  components from this law follows exactly the same law as for one component, a fact emphasized by Fisher. What would happen with a large number of observations in this case would be that larger and larger deviations would occur, the extremes increasing so rapidly that the mean will fluctuate by quantities of order 1.

For a component  $x_r$  satisfying the law

$$P(dx_r \mid H) = \frac{a_r}{\pi\{a_r^2 + (x_r - b_r)^2\}} dx_r \quad (4)$$

we have

$$\Omega_r(\kappa) = \begin{cases} e^{(b_r + i a_r)\kappa} & \{I(\kappa) > 0\}, \\ e^{(b_r - i a_r)\kappa} & \{I(\kappa) < 0\}. \end{cases} \quad (5)$$

† *The Advanced Theory of Statistics*, vol. 1, 1943, 99–103, 180–2.

For the sum of  $k$  such components

$$\Omega(\kappa) = \begin{cases} e^{(\Sigma b_r + i \Sigma a_r) \kappa} & \{I(\kappa) > 0\}, \\ e^{(\Sigma b_r - i \Sigma a_r) \kappa} & \{I(\kappa) < 0\}. \end{cases} \quad (6)$$

Hence the sum follows the law

$$P(d \sum x_r | H) = \frac{\sum a_r}{\pi \{(\sum a_r)^2 + (\sum x_r - \sum b_r)^2\}} d \sum x_r. \quad (7)$$

Thus the  $a_r$  and  $b_r$  are both additive. This can also be proved by direct integration for the combination of two components and generalized by mathematical induction.

For the Type VII law with index 2, if we reduce the scale in the ratio  $k^{-1/2}$  and combine  $k$  components, we have for the resultant

$$\Omega(\kappa) = \begin{cases} (1 - i\kappa/\sqrt{k})^k e^{i\kappa\sqrt{k}} & \{I(\kappa) > 0\}, \\ (1 + i\kappa/\sqrt{k})^k e^{-i\kappa\sqrt{k}} & \{I(\kappa) < 0\}, \end{cases} \quad (8)$$

and the probability density is

$$\begin{aligned} \frac{dF}{dx} = G = \frac{1}{2\pi i} \int_0^{i\infty} e^{-\kappa x} (1 - i\kappa/\sqrt{k})^k e^{i\kappa\sqrt{k}} d\kappa + \\ + \frac{1}{2\pi i} \int_{-i\infty}^0 e^{-\kappa x} (1 + i\kappa/\sqrt{k})^k e^{-i\kappa\sqrt{k}} d\kappa. \end{aligned} \quad (9)$$

Apart from the factor in  $x$  the integrands are real and positive and become exponentially small within a distance from the origin of order  $k^{-1/2}$ . We can approximate to the logarithm of the integrand in powers of  $k^{-1/2}$  and find

$$G = \frac{1}{2\pi i} \int_{-i\infty}^{i\infty} \exp\left\{-\kappa x + \frac{1}{2}\kappa^2 + O\left(\frac{\kappa^3}{\sqrt{k}}\right)\right\} d\kappa, \quad (10)$$

and the term in  $\kappa^3$  is negligible. Then

$$G \doteq \frac{1}{\sqrt{(2\pi)}} \exp(-\frac{1}{2}x^2). \quad (11)$$

This will be valid provided  $x$  is not comparable with  $k^{1/2}$ . If it is of order  $k^{1/2}$  or larger,  $\kappa x$  and the neglected terms in  $\kappa^3$  will be comparable. We have therefore for large  $k$  an approximation of the same nature as that found for the binomial; the normal law is a good approximation over a range that includes most of the chance.

If  $x$  is comparable with  $k^{1/2}$  or larger a different form of approximation is necessary. The method of steepest descents is also unsuitable because there is a branch-point at the origin and the paths of steepest

descent from it do not go near the saddle-points. But for the two parts the integrands fall off most rapidly in directions in the first and fourth quadrants respectively, and we can replace the integrals by those along the real axis and then apply Watson's lemma.† Then

$$G = \frac{1}{2\pi i} \int_0^{\infty} e^{-\kappa x} \{ (1 - i\kappa/\sqrt{k})^k e^{i\kappa\sqrt{k}} - (1 + i\kappa/\sqrt{k})^k e^{-i\kappa\sqrt{k}} \} d\kappa \quad (12)$$

and we want the imaginary part of the integral for  $\kappa$  small. The first non-zero term is

$$\frac{1}{3\pi} \int_0^{\infty} \frac{\kappa^3}{k^{1/2}} e^{-\kappa x} d\kappa = \frac{2}{\pi k^{1/2} x^4}. \quad (13)$$

This is proportional to 2.62(4) for  $x$  large, but it is divided by  $\sqrt{k}$ ; higher terms will involve higher powers of  $k^{-1/2}$ . The effect of combining several components is therefore to give an approach to the normal up to an indefinitely increasing multiple of the standard error; beyond this multiple the law retains the original form except that all ordinates are reduced in approximately the same ratio. The higher moments do in fact remain infinite, but the area of the tails is greatly reduced.

2.65. We can make a little further progress by considering cases where the fourth moment is finite. We shall have

$$\Omega(\kappa) = 1 + \frac{1}{2}\kappa^2 + \frac{1}{6}\mu_3\kappa^3 + \frac{1}{24}\mu_4\kappa^4 + o(\kappa^4) \quad (1)$$

and if we contract the scale in the ratio  $k^{-1/2}$  and combine  $k$  components,

$$G \doteq \frac{1}{2\pi i} \int_{-i\infty}^{i\infty} \left\{ 1 + \frac{\kappa^2}{2k} + \frac{\mu_3\kappa^3}{6k^{3/2}} + \frac{\mu_4\kappa^4}{24k^2} + o\left(\frac{\kappa^4}{k^2}\right) \right\}^k e^{-\kappa x} d\kappa. \quad (2)$$

If some higher moment is infinite, the corresponding derivative of  $\Omega(\kappa)$  will not exist at  $\kappa = 0$ , and we cannot immediately apply the method of steepest descents to (2) because the integrand is not analytic. But (2) is a valid approximation when  $\kappa = O(k^{1/2})$ , and for large  $\kappa$  the integrand is small. Hence if we drop the last term the error will be negligible, and we can apply steepest descents because without this term the integrand is analytic. Then, for large  $k$ ,

$$G \sim \frac{1}{2\pi i} \int \exp \left\{ -\kappa x + \frac{\kappa^2}{2} + \frac{\mu_3\kappa^3}{6k^{1/2}} + \frac{(\mu_4-3)\kappa^4}{24k} \right\} d\kappa \quad (3)$$

† H. and B. S. Jeffreys, *Methods of Mathematical Physics*, pp. 471, 668.



and if we take the path through  $x$  we shall have, nearly,

$$G = \frac{1}{\sqrt{(2\pi)}} \exp(-\frac{1}{2}x^2) \exp\left\{\frac{\mu_3 x^3}{6k^{1/2}} + \frac{(\mu_4 - 3)x^4}{24k}\right\}. \quad (4)$$

The correcting factor will become important if  $x$  is of order  $k^{1/6}(6/\mu_3)^{1/3}$  or  $\{24k/(\mu_4 - 3)\}^{1/4}$ , whichever is the smaller. Thus symmetry and approximate normality for the separate components will favour rapid approach to normality for the resultant. There is evidence that some errors of observation follow a Type VII law with index about 4.† For this, if  $\mu_2 = 1$ ,  $\mu_3 = 0$ ,  $\mu_4 = 5$ , and the correcting factor is  $\exp(x^4/12k)$ , for  $x$  not too large.

The conditions for the normal law to hold are fairly well satisfied in some cases, especially where the observed value is the mean of several crude readings. Thus in the standard method of determining the magnetic dip both ends of the needle are read, the needle turned over, the case rotated, and the magnetization reversed to eliminate various systematic errors. The error of the mean is then the resultant of sixteen components, presumably with the same finite second moment, and the normal law should be right up to about  $(12 \times 16)^{1/4} = 3.8$  times the standard error. In Bullard's observations of gravity in East Africa,‡ two separate swings of the pendulums in the field were compared with two in Cambridge taken at the same time; the error is therefore the resultant of four components, and if the separate laws have index 4 the normal law should hold up to about 2.6 times the standard error. But where there is a dominating source of error there may well be considerable departures from the normal law.

The normal law of error cannot therefore be theoretically proved. Its justification is that in representing many types of observations it is apparently not far wrong, and is much more convenient to handle than others that might or do represent them better. Various theoretical attempts at justification have been made, notably Gauss's proof that if the mean is the most probable value, the normal law must hold. But the argument would equally imply that since we know many cases where the law does not hold the mean is not the best estimate. Indeed, we have had Cauchy's case where the mean is no better than one observation; but with a different way of making the estimate we could get much higher accuracy from many observations than from one even with this law. Whittaker and Robinson (p. 215) give a theoretical argument for the principle of the arithmetic mean, but this is fallacious. It depends

† See later, p. 290.

‡ *Phil. Trans. A*, 235, 1936, 445-531.

on confusion between the measurement of two different quantities in terms of the same unit and of the same quantity with respect to two different units, and between the difference of two quantities with regard to the same origin and the same quantity with regard to different origins. The irrelevance of the unit and origin may be legitimate axioms, but are replaced by the former pair in the course of the argument.†

**2.66.** When several components following the same symmetrical law with a finite range are combined, the approach to the normal is very rapid. Thus an elementary law may consist of chances  $\frac{1}{2}$  at each of  $\pm 1$ . If we combine three such components the second moment for the resultant is 3, the possible values being  $-3, -1, +1, +3$ . Compare the expectations for eight observations with those corresponding to the normal law with the same second moment, supposed rounded to the nearest odd integer:

	< -4	-3	-1	+1	+3	> +4
Binomial	0	1	3	3	1	0
Normal	0.084	0.908	3.008	3.008	0.908	0.084

For four components and sixteen observations the expectations in ranges about the even numbers are as follows:

	< -5	-4	-2	0	+2	+4	> +5
Binomial	0	1	4	6	4	1	0
Normal	0.10	0.97	3.86	6.13	3.86	0.97	0.10

In neither case do the probabilities of one observation falling in a particular range differ by more than 0.012. It can be shown that if the observations were in fact derived from a binomial law with three components, and we were given only the totals by ranges to compare by the  $\chi^2$  test, used in Pearson's way, with the postulate that they are derived from the normal law, it would take about 500 observations to reveal a discrepancy.‡

If the primitive law is a rectangular one from  $-1$  to  $+1$ , we have

$$P(dx | H) = \frac{1}{2}dx \quad (-1 < x < 1) \quad (1)$$

and

$$\Omega(\kappa) = \frac{1}{2} \int_{-1}^1 e^{\kappa x} dx = \frac{1}{2\kappa} (e^{\kappa} - e^{-\kappa}). \quad (2)$$

For two components the law will be

$$P(dx | H)/dx = \frac{1}{8\pi i} \int_L \frac{1}{\kappa^2} (e^{2\kappa} - 2 + e^{-2\kappa}) e^{-\kappa x} d\kappa = \begin{cases} \frac{1}{2} - \frac{1}{4}x & (0 < x < 2), \\ \frac{1}{2} + \frac{1}{4}x & (-2 < x < 0). \end{cases} \quad (3)$$

This is known as the triangular distribution.

† *Calculus of Observations*, pp. 215-17.

‡ *Phil. Trans. A*, 237, 1938, 235.

For three components it is

$$P(dx | H)/dx = \begin{cases} \frac{1}{18}(3+x)^2 & (-3 < x < -1), \\ \frac{1}{18}(6-2x^2) & (-1 < x < 1), \\ \frac{1}{18}(3-x)^2 & (1 < x < 3). \end{cases} \quad (4)$$

The second moments for (1), (3), and (4) are  $\frac{1}{3}$ ,  $\frac{2}{3}$ , and 1. Rescaling to give unit second moment in each case we have from (1) and (3)

$$P(dx | H) = \frac{1}{2\sqrt{3}} dx \quad (-\sqrt{3} < x < \sqrt{3}), \quad (5)$$

$$\frac{P(dx | H)}{dx} = \begin{cases} \frac{1}{\sqrt{6}} \left(1 - \frac{x}{\sqrt{6}}\right) & (0 < x < \sqrt{6}), \\ \frac{1}{\sqrt{6}} \left(1 + \frac{x}{\sqrt{6}}\right) & (-\sqrt{6} < x < 0). \end{cases} \quad (6)$$

while (4) needs no change.

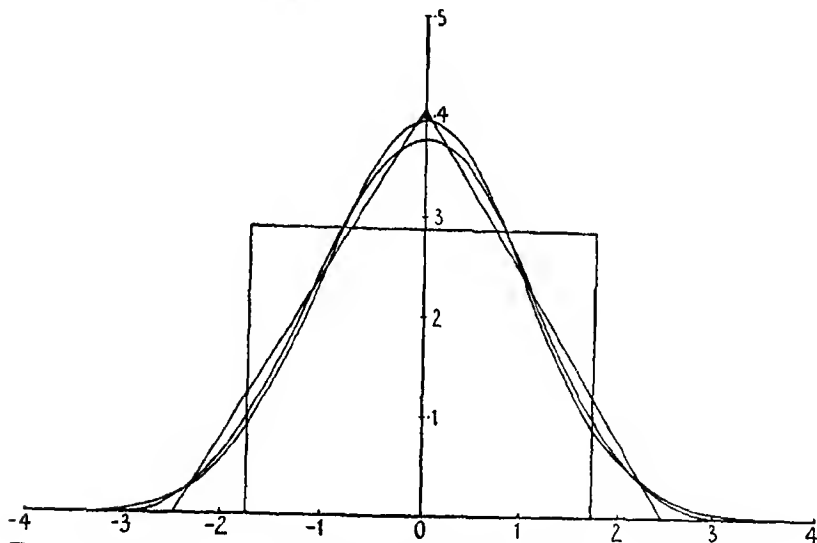


FIG. 1. Laws of equal (unit) second moment obtained by combining 1, 2, 3,  $\infty$  rectangular distributions.

We see at a glance from Fig. 1 that (6) already gives a fair approach to the normal, though it has combined only two rectangular distributions; while (4) is very close, even at the tails.

The approach to the normal is much less rapid if the component laws are asymmetrical. Thus if three components each give chances  $\frac{2}{3}$  of  $-\frac{1}{3}$  and  $\frac{1}{3}$  of  $+\frac{2}{3}$ , the expectations from the results of 27 observations are

-1	0	+1	+2
8	12	6	1

and plainly no normal law can fit all the chances within a little under 0.04.

**2.7. The  $\chi^2$  distribution.** Suppose that we have  $n$  independent variables with normal distributions of chance about zero, so that we can write

$$P(dx_1 dx_2 \dots dx_n | H) = \frac{1}{(2\pi)^{1/2n} \sigma_1 \sigma_2 \dots \sigma_n} \exp\left\{-\frac{1}{2}\left(\frac{x_1^2}{\sigma_1^2} + \frac{x_2^2}{\sigma_2^2} + \dots + \frac{x_n^2}{\sigma_n^2}\right)\right\} dx_1 dx_2 \dots dx_n. \quad (1)$$

Consider the total chance that the function

$$\chi^2 = \frac{x_1^2}{\sigma_1^2} + \frac{x_2^2}{\sigma_2^2} + \dots + \frac{x_n^2}{\sigma_n^2} \quad (2)$$

may fall in a given range. This can be got by integrating over all values of  $x_1$  to  $x_n$  that correspond to  $\chi^2$  in this range. First put

$$x_1 = \sigma_1 y_1, \quad x_2 = \sigma_2 y_2, \quad \text{etc.}$$

Then 
$$\chi^2 = \sum y^2 \quad (3)$$

and 
$$P(d\chi^2 | H) = (2\pi)^{-1/2n} \int \dots \int \exp(-\frac{1}{2}\chi^2) dy_1 \dots dy_n. \quad (4)$$

If we like we can regard the  $y$ 's as Cartesian coordinates in  $n$  dimensions and the integral with regard to them as a volume integral. But in any case in a range between two neighbouring values of  $\chi$  we can neglect the variation of  $\chi$ , while all the  $y$ 's are proportional to  $\chi$ . The integral from 0 up to a given  $\chi$ , omitting the factor  $\exp(-\frac{1}{2}\chi^2)$ , would be proportional to  $\chi^n$ ; hence the change in it due to a change in  $\chi$  is proportional to  $\chi^{n-1} d\chi$ , and now, since we can neglect the variation of  $\exp(-\frac{1}{2}\chi^2)$  in the shell, we have

$$P(d\chi^2 | H) \propto \chi^{n-1} \exp(-\frac{1}{2}\chi^2) d\chi. \quad (5)$$

The constant factor can be found by using the condition that  $\chi^2$  is certain to lie between 0 and  $\infty$ , or the Dirichlet integral may be used.

Then 
$$P(d\chi^2 | H) = \frac{1}{2^{1/2}(n-2)(\frac{1}{2}n-1)!} \chi^{n-1} \exp(-\frac{1}{2}\chi^2) d\chi. \quad (6)$$

It is easy to verify that the expectation of  $\chi^2$  is  $n$ , as is obvious from its definition. The maximum of the integrand is near  $\chi^2 = n$ . If we neglect a factor  $\chi^{-1}$  and take logarithms,

$$\frac{d^2}{d\chi^2} (n \log \chi - \frac{1}{2}\chi^2) = -2 \quad (7)$$

near the maximum, whence if  $n$  is large  $P(d\chi^2 | H)$  is nearly proportional

to  $\exp\{-(\chi - \sqrt{n})^2\}d\chi$  or to  $\exp\{-(\chi^2 - n)^2/4n\}d\chi^2$ . Thus, roughly, we can write

$$\chi^2 = n \pm \sqrt{(2n)} \quad (8)$$

as a summary expression of the rule. Tables giving  $P(\chi^2)$ , the chance that  $\chi^2$  will exceed a given value, are given by Pearson, Fisher, and Yule and Kendall.

The interest of this rule is that it often enables us to see very easily whether a set of data are consistent with a hypothesis. It is required that we shall have a set of estimates, obtained independently, on a hypothesis that gives estimates of the standard errors, and that we compare them with a set of values predicted by the hypothesis. In general the observed and theoretical values will differ by quantities of the order of the standard errors, but if we form  $\chi^2$  we have a quantity that would be increased either by an unexpected systematic variation (the random variation remaining the same), by the actual random variation being larger than that expected, or by some internal correlation that makes errors tend to repeat themselves, when the means will vary more than expected. If  $\chi^2$  is less than  $n + \sqrt{(2n)}$  we can usually say at once that the observations agree with the theory as well as could be expected, and if it is less than  $n + 2\sqrt{(2n)}$  there is no immediate need to discard the hypothesis. The matter will be treated in more detail later, but these simple considerations so often cover all that is wanted that they may as well be stated at the outset.

**2.71.** It often (or rather usually) happens that the hypothesis investigated contains some adjustable parameters, and that these are determined in such a way as to make  $\chi^2$  a minimum. If they are fewer than the  $x$ 's there will still be an outstanding variation, but we should naturally expect it to be smaller than the original one. Instead of all the  $x$ 's being independent, we must now suppose that the information with respect to them can be written

$$x_r = l_r \alpha \pm \sigma_r, \quad (9)$$

where the  $l_r$  are known, but  $\alpha$  is to be found, and  $x_r - l_r \alpha$  can be considered random. Then on this hypothesis

$$P(dx_1 \dots dx_n | \alpha H) = \frac{(2\pi)^{-1/2n}}{\sigma_1 \sigma_2 \dots \sigma_n} \exp\left\{-\sum \frac{(x_r - l_r \alpha)^2}{2\sigma_r^2}\right\} dx_1 \dots dx_n. \quad (10)$$

Now suppose that we determine the value of  $\alpha$ ,  $a$  say, that makes  $\sum (x_r - l_r \alpha)^2 / \sigma_r^2$  a minimum. Then

$$\sum \frac{l_r (x_r - l_r a)}{\sigma_r^2} = 0, \quad (11)$$

and 
$$\sum \frac{(x_r - l_r \alpha)^2}{2\sigma_r^2} = \sum \frac{(x_r - l_r a)^2}{2\sigma_r^2} + (\alpha - a)^2 \sum \frac{l_r^2}{2\sigma_r^2}. \quad (12)$$

The first term on the right is the value of  $\frac{1}{2}\chi^2$  that would be found by comparing the  $x_r$  with  $l_r a$  instead of with 0 or  $l_r \alpha$ . Hence

$$P(dx_1 \dots dx_n | \alpha H) = \frac{(2\pi)^{-1/2n}}{\sigma_1 \sigma_2 \dots \sigma_n} \exp \left\{ -\frac{1}{2}\chi^2 - (\alpha - a)^2 \sum \frac{l_r^2}{2\sigma_r^2} \right\} dx_1 \dots dx_n. \quad (13)$$

The form of this shows that the information about the  $x_r$  can be regarded as composed of three independent parts. For they would all be determined by  $a$ ,  $\chi$ , and  $n-2$  direction parameters of the form  $m_r = (x_r - l_r a)/\sigma_r \chi$ . If we change to these as new variables the three groups of chances will be independent, and by applying Theorem 12 we have

$$P(d\chi | \alpha, a, m_r, H) \propto \exp(-\frac{1}{2}\chi^2) \frac{\partial(x_1, x_2, \dots, x_n)}{\partial(a, \chi, m_1, \dots, m_{n-2})} d\chi \propto \chi^{n-2} \exp(-\frac{1}{2}\chi^2) d\chi. \quad (14)$$

Thus the determination and elimination of each adjustable constant reduces the index of  $\chi$  in the chance for the outstanding variation by 1. The difference between the number of separate data and the number of parameters allowed for is usually called the *number of degrees of freedom*. If this is identified with the  $n$  of (6) the formula will always hold.

It may be noticed that on the left  $d\chi$  means the proposition that  $\chi$  will lie in a *particular* range  $d\chi$ ;  $d\chi^2$  means that  $\chi^2$  will lie in the corresponding range  $d\chi^2$ . These propositions are equivalent and can therefore be interchanged in the expression on the left by Theorem 3.

**2.72.** If there is a linear constraint on the data, so that

$$\sum m_r x_r = 0,$$

this also will remove one variable from the integration and reduce the degrees of freedom by 1 and also the index of the distribution.

**2.73.**  $\chi^2$  was first obtained by Pearson in relation to a problem of sampling.† In the latter case it can be simply derived from the last remark. Suppose that we are sampling an enormous population of several different types, and that the expectations in a sample, given the time of sampling and the proportions in the population, are  $m_1, m_2, \dots, m_p$ . Then if these are moderate numbers and the occurrences of members of different types do not interfere, each type will give an

† *Phil. Mag.* 50, 1900, 157-75.

independent Poisson distribution and the expected number may be written  $m_r \pm \sqrt{m_r}$ . If the observed numbers are  $n_r$  we have therefore

$$\chi^2 = \sum (n_r - m_r)^2 / m_r$$

taken over all types. The degrees of freedom will be  $p$ . Such a case might be realized if we were observing a phenomenon for a finite time, so that the total number of events was subject to a sampling variation, besides the separate variations of the numbers of the types.

**2.74.** But if we are extracting from a population a sample of given size, the total number of the sample is known as  $N = \sum n_r$ . If the expectations are assessed in given ratios, but now are subject to the total of  $n_r$  being  $N$ , we have introduced a linear constraint and the number of degrees of freedom will be  $p-1$ . A detailed treatment, following Pearson, is as follows. We return to the multinomial rule. If  $N$  is prescribed, and subject to  $N$  the expectations are  $m_1, \dots, m_p$ , the probability of a sample  $n_1, \dots, n_p$  is

$$P(n_1, n_2, \dots, n_p | NH) = \frac{N!}{n_1! n_2! \dots n_p!} \left(\frac{m_1}{N}\right)^{n_1} \left(\frac{m_2}{N}\right)^{n_2} \dots \quad (1)$$

Put

$$n_r = m_r + \alpha_r N^{1/2}, \quad (2)$$

where  $\sum \alpha_r = 0$ . Then

$$\log \prod (n_r!) \doteq \frac{1}{2} p \log 2\pi - \sum n_r + \sum (n_r + \frac{1}{2}) \log n_r, \quad (3)$$

$$\log N! \doteq \frac{1}{2} \log 2\pi - N + (N + \frac{1}{2}) \log N, \quad (4)$$

$$\begin{aligned} P(n_1, n_2, \dots, n_p | NH) &\doteq \frac{N^{N+1/2}}{(2\pi)^{1/2(p-1)} \prod (n_r^{n_r+1/2})} \frac{\prod m_r^{n_r}}{N^N} \\ &= \frac{N^{1/2}}{(2\pi)^{1/2(p-1)} \prod (n_r^{1/2}) \prod (1 + \alpha_r N^{1/2} / m_r)^{m_r + \alpha_r N^{1/2}}}, \end{aligned} \quad (5)$$

which gives, on approximating to order  $\alpha_r^2$ ,

$$\frac{N^{1/2}}{(2\pi)^{1/2(p-1)} \prod (m_r^{1/2})} \exp\left(-\frac{1}{2} \sum \frac{N \alpha_r^2}{m_r}\right) \quad (6)$$

and 
$$\sum \frac{N \alpha_r^2}{m_r} = \sum \frac{(n_r - m_r)^2}{m_r} = \chi^2. \quad (7)$$

The probability distribution of  $\chi$ , given the  $m_r$ , is now to be found by integration. But only  $p-1$  of the  $n_r$  can be varied independently, and the result will be

$$P(d\chi | m_1 \dots m_p H) \propto \chi^{p-2} \exp(-\frac{1}{2} \chi^2) d\chi. \quad (8)$$

**2.75.** If the analysis refers to a rectangular contingency table and we wish to test whether the elements agree with the hypothesis that the chances in different rows are in proportion, further degrees of freedom

disappear. For in such a case the ratios of the total chances in the rows or in the columns are not fixed initially and must be estimated from the data. Thus if there are  $m$  rows and  $n$  columns, we fix  $m$  parameters from the numbers in the rows and  $n-1$  from the columns. The expectations being made in proportion, consistently with the row and column totals, the number of degrees of freedom that remain in  $\chi^2$  is

$$mn - m - (n-1) = (m-1)(n-1).$$

If  $m = n = 2$  the number therefore reduces to 1.

2.76. The  $\chi^2$  analysis is of enormous use. It is easy to apply, and very often is enough to answer the question asked. This means really that the hypothesis stated is very often right and the predictions made by it come off. It does not, however, always go into sufficient detail. More will be said about this under significance tests. The trouble is that it combines all degrees of freedom together as if they were all relevant to the same question, whereas only part of the information in them may be relevant. If, for instance, we have a set of data with 32 degrees of freedom, the expected  $\chi^2$  on the hypothesis of complete randomness will be  $32 \pm 8$ , which means that in the ordinary course of events it may be anything from 24 to 40 and might go beyond this range without anything but random error being involved. If there is actually a systematic variation whose amount is four times its standard error, it will contribute 16 to  $\chi^2$ ; but if the other degrees of freedom happen to contribute only 24 the total will still be 40, which would pass as entirely random. But a systematic variation of 4 times its standard error would be accepted as genuine by any significance test if it was tested directly. The trouble is that with regard to a large number of data we may want to ask several questions. To some of them the answer will be 'yes', to others 'no'. But if we try to sum up all the information in one number we shall not know what question we have answered. It is desirable to arrange the work, when several questions arise simultaneously, so as to provide answers to each of them separately. When this is done it is still found that the  $\chi^2$  form persists, but it is now broken up into separate parts each of which has its own message.

The passage from (5) to (6) above involves the neglect of cubic terms. In Pearson's earlier work he ignored the resulting errors, sometimes applying the result when the expectation was considerably less than 1. Later he recommended grouping the small expectations together so that the expectation in no group would be less than 5. This has the disadvantage that in, for instance, a test of the normal law of errors,



an observation in a range where there might be a 0.001 chance that any would occur on the normal law, and taken by itself would be strong evidence against the law, cannot be considered except in combination with several others, and there is considerable loss in sensitiveness. Both methods have drawbacks in dealing with small groups, but where the expectations are over 1 the earlier method seems to be the better; where they are under 1 the only solution seems to be to introduce a new parameter explicitly and estimate it. Then the relevant part of  $\chi^2$  is the square of the ratio of the new parameter to its standard error.

**2.8. The  $t$  and  $z$  distributions.** Suppose that we have  $n$  observations derived from the normal law with true value  $x$  and standard error  $\sigma$ . Their joint chance is

$$P(dx_1 \dots dx_n | x, \sigma, H) = \frac{1}{(2\pi)^{1/2n} \sigma^n} \exp \left\{ -\frac{1}{2\sigma^2} \sum (x_r - x)^2 \right\} dx_1 \dots dx_n. \quad (1)$$

Put  $n\bar{x} = \sum x_r; \quad (n-1)s^2 = ns'^2 = \sum (x_r - \bar{x})^2. \quad (2)$

Then  $\bar{x}$  is the arithmetic mean and  $s$  is the standard deviation as usually defined. We shall call  $s'$  the mean square deviation.  $x$ ,  $s$ , and  $s'$  are all determinate functions of the observed values. In the present problem writing is simplified by using  $s'$  rather than  $s$ , but when we come to the method of least squares we shall find that  $s$  has advantages. Also

$$\begin{aligned} \sum (x_r - x)^2 &= \sum \{(x_r - \bar{x}) + (\bar{x} - x)\}^2 \\ &= \sum (x_r - \bar{x})^2 + 2(\bar{x} - x) \sum (x_r - \bar{x}) + n(\bar{x} - x)^2. \end{aligned} \quad (3)$$

The second term vanishes by the definition of  $\bar{x}$ , and the result is

$$ns'^2 + n(\bar{x} - x)^2.$$

Hence

$$P(dx_1 \dots dx_n | x, \sigma, H) = \frac{1}{(2\pi)^{1/2n} \sigma^n} \exp \left[ -\frac{n}{2\sigma^2} \{(\bar{x} - x)^2 + s'^2\} \right] dx_1 \dots dx_n. \quad (4)$$

Thus  $\bar{x}$  and  $s$  or  $s'$  are what Fisher calls *sufficient statistics*. A 'statistic' in his terminology is any function of the observations that we might choose to provide an estimate of an unknown parameter in a law. We have seen that, whatever the prior probability may be, the observations enter into the posterior probability only through the likelihood, which in this case is the function we have just given. Also in practice the observations are not exact determinations since we read only to the nearest convenient multiple of some convenient unit. A reading of 15.3 mm. means really an observation between 15.25 and 15.35 mm.

This range of 0.1 mm. would replace  $dx_r$  in practice, and it is the same whatever the parameters in the law. Hence, when we apply the principle of inverse probability, the factor  $dx_1 \dots dx_n$  is the same for all values of the unknowns  $x$  and  $\sigma$ , and will cancel. It follows that the whole of the information with respect to  $x$  and  $\sigma$  that is contained in the observations is summarized in the two statistics  $\bar{x}$  and  $s$ . When this occurs it is unnecessary to make further reference to the observations apart from these statistics, which are therefore called *sufficient*. A definition of a sufficient statistic is as follows. Whenever the likelihood, apart from factors independent of the unknown parameters to be estimated, can be expressed as a function of the unknown parameters, the number of observations, and a number of functions of the observations equal to the number of unknown parameters, those functions of the observations are called sufficient statistics.

For various purposes we require to know the joint probability distribution of  $\bar{x}$  and  $s'$ , given  $x$  and  $\sigma$ . Then we must consider a pair of ranges of  $\bar{x}$  and  $s'$  and form the integral of (4) over all values of the observable values  $x_r$  that give  $\bar{x}$  and  $s'$  in these ranges. This is easily done as follows, by translating into analytic language a geometrical argument due to Fisher. We can regard  $x_r$  as a set of rectangular coordinates of a point in  $n$ -dimensional space, and then  $\sum (x_r - x)^2$  is the square of the distance of this point from a point all of whose coordinates are  $x$ . But we can rotate the axes in any way, and this will still hold for the new axes. In analytic language, we can form  $n$  linear functions of the  $x_r$  such that if a new function is  $x'_i$

$$x'_i = \sum_r a_{ir} x_r, \quad (5)$$

$$\text{where} \quad \sum_i a_{ir}^2 = 1; \quad \sum_r a_{ir}^2 = 1; \quad \sum_r a_{ir} a_{jr} = 0, \quad (6)$$

and this can be done in an infinity of ways. We can choose one of the  $x'_i$  to be

$$x'_1 = \sum_r x_r / \sqrt{n} = \bar{x} \sqrt{n}. \quad (7)$$

Applying this to the point  $(x, x, x, \dots)$  gives  $(x\sqrt{n}, 0, 0, \dots)$ .

Then

$$\begin{aligned} & \iiint \dots \int \exp \left\{ -\frac{1}{2\sigma^2} \sum (x_r - x)^2 \right\} dx_1 \dots dx_n \\ &= \iiint \dots \int \exp \left\{ -\frac{1}{2\sigma^2} (x'_1 - x\sqrt{n})^2 - \frac{1}{2\sigma^2} \sum' x_i'^2 \right\} dx'_1 \dots dx'_n \quad (8) \end{aligned}$$

through any region, where  $\sum'$  denotes summation for all  $i$  except  $i = 1$ .

Also

$$\sum' x_i'^2 = \sum (x_r - \bar{x})^2 = ns'^2. \quad (9)$$

Hence if we consider a region between two fixed values of  $x'_1$  and two fixed values of  $s'$ , the integral breaks up into two factors

$$I_1 = \int_{x'_1}^{x'_1 + dx'_1} \exp\left\{-\frac{1}{2\sigma^2}(x'_1 - x\sqrt{n})^2\right\} dx'_1, \quad (10)$$

$$I_2 = \iint \dots \int \exp\left\{-\frac{ns'^2}{2\sigma^2}\right\} dx'_2 \dots dx'_n, \quad (11)$$

integration in the latter case being over all values such that

$$ns'^2 \leq \sum x'_i{}^2 \leq n(s' + ds')^2. \quad (12)$$

Within short ranges of  $x'_1$  and  $s'$ , therefore, the integral

$$\propto \exp\left\{-\frac{1}{2\sigma^2}(x'_1 - x\sqrt{n})^2\right\} dx'_1 \cdot s'^{n-2} \exp\left\{-\frac{ns'^2}{2\sigma^2}\right\} ds' \quad (13)$$

$$\propto \exp\left\{-\frac{n}{2\sigma^2}(\bar{x} - x)^2\right\} d\bar{x} \cdot s'^{n-2} \exp\left\{-\frac{ns'^2}{2\sigma^2}\right\} ds'. \quad (14)$$

The constant factor is determined by the condition that  $\bar{x}$  is certain to be between  $\pm\infty$  and  $s'$  between 0 and  $\infty$ . Hence

$$P(d\bar{x}ds' | x, \sigma, H) = \sqrt{\left(\frac{n}{2\pi}\right)} \frac{1}{\sigma} \exp\left\{-\frac{n}{2\sigma^2}(\bar{x} - x)^2\right\} d\bar{x} \cdot \frac{n^{1/2n-1/2} s'^{n-2}}{2^{1/2(n-3)} (\frac{1}{2}n - \frac{3}{2})! \sigma^{n-1}} \exp\left\{-\frac{ns'^2}{2\sigma^2}\right\} ds'. \quad (15)$$

The argument fails if  $n = 1$ , for then  $s'$  is necessarily 0 and the factor  $I_2$  does not arise.

$$\text{Now put} \quad \bar{x} - x = s'z \quad (16)$$

and transform to variables  $s'$  and  $z$ . We have now

$$P(dzds' | x, \sigma, H) = \frac{n^{1/2n}}{\sqrt{\pi} \cdot 2^{1/2n-1} (\frac{1}{2}n - \frac{3}{2})!} \frac{s'^{n-1}}{\sigma^n} \exp\left\{-\frac{ns'^2}{2\sigma^2}(1+z^2)\right\} ds' dz, \quad (17)$$

and finally, on integrating with regard to  $s'$ ,

$$P(dz | x, \sigma, H) = \frac{(\frac{1}{2}n - 1)!}{\sqrt{\pi} \cdot (\frac{1}{2}n - \frac{3}{2})!} (1+z^2)^{-1/2n} dz. \quad (18)$$

This rule was first obtained by W. L. Gosset, a prominent statistical writer who used the nom de plume of 'Student'.† Its remarkable feature is that it is independent of  $x$  and  $\sigma$ , which may therefore be suppressed; their actual values are irrelevant to  $z$ , and their existence is implied by  $H$ , which includes the statement that the normal law holds in the

† *Biometrika*, 6, 1908, 1-25.

problem under discussion. It may be transformed by introducing the quantities

$$s_x = \frac{s'}{(n-1)^{1/2}} = \frac{s}{n^{1/2}} = \left\{ \frac{\sum (x_r - \bar{x})^2}{n(n-1)} \right\}^{1/2}, \quad (19)$$

$$t = \frac{\bar{x} - x}{s_x} = (n-1)^{1/2} z. \quad (20)$$

Then  $s_x$  is the usual conventional estimate of the standard error of a mean† and  $t$  is the ratio of the actual error of the mean to the estimated standard error. We shall then have

$$P(dt | x, \sigma, H) = P(dt | H) = \frac{(\frac{1}{2}n-1)!}{\sqrt{\pi} \cdot (n-1)^{1/2} (\frac{1}{2}n - \frac{3}{2})!} \left(1 + \frac{t^2}{n-1}\right)^{-1/2n} dt. \quad (21)$$

This is now the usually adopted form, and is called the  $t$  distribution. If  $n$  is large it tends to the normal with standard error 1, but for moderate values of  $n$  it is more widely spread to large values of  $t$ . This represents the fact that, given  $x$  and  $\sigma$ , the probabilities of different values of  $\bar{x}$  and  $s'$  are independent. Consequently, while those of  $\bar{x}$  follow the normal law with standard error  $\sigma/\sqrt{n}$ , in any individual case the error of  $x$  may be associated with a value of  $s'$  either more or less than  $\sigma$ , and  $s_x$  as calculated from  $s$  may be either more or less than  $\sigma/\sqrt{n}$ . The result is that there is a considerable chance that an error of  $x$  larger than  $\sigma/\sqrt{n}$  will be associated with a value of  $s_x$  less than  $\sigma/\sqrt{n}$ , and the result will be to give an excess chance of large values of  $t$  in comparison with that for  $x/\sigma$  on the normal law.

**2.81.** Suppose now that we have two separate samples of  $n_1$  and  $n_2$  derived from a normal law with the same parameters, and that their means and mean square deviations are  $\bar{x}_1$  and  $\bar{x}_2$ ,  $s'_1$  and  $s'_2$ . What is the joint chance of these four quantities lying in prescribed ranges, given  $x$  and  $\sigma$ ? Since the law is one of chance neither set can give any information about the other when  $x$  and  $\sigma$  are given; hence by the product rule

$$\begin{aligned} P(d\bar{x}_1 d\bar{x}_2 ds'_1 ds'_2 | x, \sigma, H) \\ = \frac{(n_1 n_2)^{1/2}}{2\pi\sigma^2} \exp\left\{-\frac{n_1}{2\sigma^2}(\bar{x}_1 - x)^2\right\} d\bar{x}_1 \exp\left\{-\frac{n_2}{2\sigma^2}(\bar{x}_2 - x)^2\right\} d\bar{x}_2 \times \\ \times \frac{n_1^{1/2} n_1^{-1/2} s'^{n_1-2}_1}{2^{1/2}(n_1-3)(\frac{1}{2}n_1 - \frac{3}{2})! \sigma^{n_1-1}} \exp\left\{-\frac{n_1 s'^2_1}{2\sigma^2}\right\} ds'_1 \times \\ \times \frac{n_2^{1/2} n_2^{-1/2} s'^{n_2-2}_2}{2^{1/2}(n_2-3)(\frac{1}{2}n_2 - \frac{3}{2})! \sigma^{n_2-1}} \exp\left\{-\frac{n_2 s'^2_2}{2\sigma^2}\right\} ds'_2, \end{aligned} \quad (22)$$

† It has no unique standard error since the posterior probability of the true value, given the mean and standard deviation, is not normally distributed.

which is the product of four independent factors. Now consider the chance that  $s'_1$  will lie between  $s'_2 y$  and  $s'_2(y+dy)$ . For all values of  $\bar{x}_1$  and  $\bar{x}_2$  we have, by Theorem 12,

$$P(dy ds'_2 | \bar{x}_1, \bar{x}_2, x, \sigma, H) = \frac{n_1^{1/2} n_1^{-1/2} n_2^{1/2} n_2^{-1/2} s'_2{}^{n_1+n_2-3} y^{n_1-2}}{2^{1/2(n_1+n_2-6)} (\frac{1}{2}n_1 - \frac{3}{2})! (\frac{1}{2}n_2 - \frac{3}{2})! \sigma^{n_1+n_2-2}} \exp\left(-\frac{n_2+n_1 y^2}{2\sigma^2} s'_2{}^2\right) dy ds'_2 \quad (23)$$

and, integrating with regard to  $s'_2$ ,

$$P(dy | \bar{x}_1, \bar{x}_2, x, \sigma, H) = \frac{2n_1^{1/2} n_1^{-1/2} n_2^{1/2} n_2^{-1/2} (\frac{1}{2}n_1 + \frac{1}{2}n_2 - 2)! y^{n_1-2}}{(\frac{1}{2}n_1 - \frac{3}{2})! (\frac{1}{2}n_2 - \frac{3}{2})! (n_2 + n_1 y^2)^{1/2(n_1+n_2-2)}} dy. \quad (24)$$

Now put  $y = e^z$ .

$$P(dZ | \bar{x}_1, \bar{x}_2, x, \sigma, H) = \frac{2n_1^{1/2} n_1^{-1/2} n_2^{1/2} n_2^{-1/2} (\frac{1}{2}n_1 + \frac{1}{2}n_2 - 2)!}{(\frac{1}{2}n_1 - \frac{3}{2})! (\frac{1}{2}n_2 - \frac{3}{2})!} \frac{e^{(n_1-1)Z} dZ}{(n_2 + n_1 e^{2Z})^{1/2(n_1+n_2-2)}}. \quad (25)$$

This, with a change of variable, is Fisher's  $z$  distribution.† If we take  $\nu_1 = n_1 - 1$ ,  $\nu_2 = n_2 - 1$  (following Yule and Kendall in this notation), we have  $\nu_1 s_1^2 = n_1 s_1'^2$ ,  $\nu_2 s_2^2 = n_2 s_2'^2$ ,  $\log(s_1/s_2) = z$ ,

$$P(dz | \bar{x}_1, \bar{x}_2, x, \sigma, H) = \frac{2\nu_1^{1/2} \nu_1 \nu_2^{1/2} (\frac{1}{2}\nu_1 + \frac{1}{2}\nu_2 - 1)!}{(\frac{1}{2}\nu_1 - 1)! (\frac{1}{2}\nu_2 - 1)!} \frac{e^{\nu_1 z} dz}{(\nu_2 + \nu_1 e^{2z})^{1/2(\nu_1 + \nu_2)}}. \quad (26)$$

This is Fisher's form. It is curious that the factors that arise in the transformation should cancel so completely. In practice there is an arbitrariness as to which of the standard deviations we should call  $s_1$ ; the larger is taken, so that  $z$  in actual use is always positive. It is easy to verify that interchanging  $s_1$  and  $s_2$  and reversing the sign of  $z$  leaves (26) unaltered. But apart from this conventional restriction  $z$  can range from  $-\infty$  to  $+\infty$ , unlike  $y$ , which can only range from 0 to  $\infty$ , and the law for  $z$  is therefore much more symmetrical. The law is in fact nearly normal for moderate departures of  $z$  from 0, and may be conveniently represented by

$$z = 0 \pm \left\{ \frac{1}{2} \left( \frac{1}{\nu_1} + \frac{1}{\nu_2} \right) \right\}^{1/2}.$$

Detailed tables of the values of  $z$  with 5, 1 and 0.1 per cent. chances of being exceeded on the hypothesis of random variation are given by Fisher.‡

**2.82.** The  $z$  rule may be regarded as a generalization of  $\chi^2$ . The  $\chi^2$  rule assumes that the data are either derived from the normal law with

† *Proc. Roy. Soc. A*, **121**, 1928, 669.

‡ *Statistical Methods for Research Workers*, Table VI.

known standard errors, or approximately so derived with standard errors calculable from frequencies, and the probable scatter of the data is compared with the known standard errors. In the  $z$  rule, the scatter of one set of estimates is compared with that of another set, each being measured by the standard deviation and not by the standard error, and consequently both numbers of degrees of freedom appear in the result. But it is supposed that each estimate of either set has the same standard error. This is achieved in biological experiments by what is called a balanced design (cf. 4.9). In physics it is hardly ever achieved; the essence of comparison of physical estimates is usually that they have been obtained by different methods and consequently have different standard errors. We therefore need a method to replace the  $z$  rule in such conditions; we can hope only for an approximate answer, but some answer is necessary.

If we have several series of estimates  $x_r$  with estimated standard errors  $c_r$  based on  $\nu_r$  d.f., we might suggest forming the sum

$$\sum \frac{x_r^2}{c_r^2} = \sum t_r^2 \quad (1)$$

for the series together, measuring each  $x_r$  from a weighted mean of the  $x_r$ . This is the simplest analogue of  $\chi^2$ . When all the  $\nu_r$  are large it is fairly satisfactory. If there are  $n$  estimates the number of degrees of freedom is  $n-1$ . But if the  $\nu_r$  are not large this function will not follow the same rule as  $\chi^2$ . The expectation of  $t_r^2$  from (1) is not 1 but  $\nu_r/(\nu_r-2)$  for  $\nu_r > 2$ ; for  $\nu_r \leq 2$  it is infinite. Consequently, if we estimate  $\chi^2$ , using the estimated standard errors, the estimate will be about

$$\sum \frac{\nu_r}{\nu_r-2} - 1 \quad (2)$$

instead of  $n-1$ . This may be serious. Suppose that we have 10 series of 5 observations each, and form  $\chi^2$  in this way from the means. The expectation of  $\chi^2$  will be 19 instead of 9. But on 9 d.f.  $\chi^2 = 19$  is nearly up to the 2 per cent. point, and such a set of means will habitually be judged discordant even if the variation is wholly random.

A better method is suggested by the central limit theorem. We have, if  $E$  denotes an expectation,

$$E(t^2 - Et^2)^2 = Et^4 - (Et^2)^2 = \frac{2\nu^2(\nu-1)}{(\nu-2)^2(\nu-4)}, \quad (3)$$

and if 
$$t'^2 = t^2 \frac{\nu-2}{\nu} \sqrt{\frac{\nu-4}{\nu-1}} \quad (4)$$

the expectation of  $(t'^2 - Et'^2)^2$  is always 2 for  $\nu > 4$ , and  $\sum (t_r'^2 - Et_r'^2)$

will have a nearly normal probability distribution for  $n$  more than about 3 or 4. Then

$$Et_r'^2 = \sqrt{\frac{\nu_r - 4}{\nu_r - 1}}, \quad \sum t_r'^2 = \sum \sqrt{\frac{\nu_r - 4}{\nu_r - 1}} \pm \sqrt{(2n)} \quad (5)$$

if the true value is taken as 0. If one weighted mean is determined it will be allowed for approximately by multiplying the first term by  $(n-1)/n$  and replacing  $\sqrt{(2n)}$  by  $\sqrt{(2n-2)}$  in the second. It now becomes impossible to include in the test any estimates based on fewer than 4 d.f., but if those with  $\nu_r > 4$  are found accordant they can be combined, and then those with  $\nu_r \leq 4$  can be compared with them individually.

The method is necessarily rough, but should serve as a useful compromise capable of being used in the same way as  $\chi^2$ . Like  $\chi^2$  and  $z$ , it will not always be the end of the matter, but will provide a simple way of seeing whether it is worth while to go into greater detail.

### III

#### ESTIMATION PROBLEMS

‘We’ve got to stand on our heads, as men of intellect should.’

R. AUSTIN FREEMAN, *The Red Thumb Mark*

**3.0.** IN the problems of the last chapter we were considering the probabilities that various observable events would occur, given certain laws and the values of all parameters included in these laws. The usual use of these results is that they provide the likelihood for different values of the parameters; then, taking the observed results as given, and using the principle of inverse probability, we can assess the relative probabilities of the different values of the parameters. A problem of estimation is one where we are given the form of the law, in which certain parameters can be treated as unknown, no special consideration needing to be given to any particular values, and we want the probability distributions of these parameters, given the observations.

Now from any finite number of observations we can never evaluate more than a certain number of parameters. A sample  $(l, m)$  cannot determine more than two parameters and, since  $l+m$  is in practice chosen for convenience and has no reference beyond the sample, there will be only one parameter that has any relevance beyond the sample itself. A set of  $n$  quantitative observations cannot determine more than  $n$  adjustable parameters; but if we always admitted the full  $n$  we should be back at our original position, since a new parameter would imply a new function, and we should change our law with every observation. Thus the principle that laws have some validity beyond the original data would be abandoned. It is necessary, therefore, to the statement of a scientific law that it involves a number of adjustable parameters (possibly none) and that new observations do not alter the form of the law, though they may alter the estimates of the parameters. The likelihood of a given set of observations has no definite value unless the form of the law is given and all the parameters in the law are explicitly stated.

On the other hand, a law is not a final statement. By rule 5 we can rule out no law as impossible *a priori*, and if a true law involves  $n$  parameters it could not be found until there are more than  $n$  relevant observations. Hence the number of parameters in the laws that it is possible to consider at any time depends on the number of observations. Thus it is a necessity of progress that laws must be considered, on the whole, in the order of increasing number of adjustable parameters.



The function of significance tests is to provide a way of arriving, in suitable cases, at a decision that at least one new parameter is needed to give an adequate representation of the existing data and valid inferences to future ones. But we must not deny in advance that those already considered are adequate, the outstanding variation being legitimately treated as random. Though we do not claim that our laws are necessarily final statements, we claim that they may be, and that on sufficient evidence they have high probabilities. But by rule 5 we can set no limit to the number of possible laws, and this is the same as saying that the number is infinite. If all laws had the same prior probability it would be infinitesimal, and would remain infinitesimal on any amount of evidence. Thus there could be no stop, nor even a temporary pause, unless we agree that every law has a finite prior probability. But then if there are an infinite number of possible laws their prior probabilities must form a convergent series.

This result implies the possibility of arranging possible laws in an order of decreasing prior probability. What can this order be? The methods capable of being adopted, which are mainly those already in use, provide our answer. It is the order in which the laws ordinarily arise for consideration, that of increasing number of adjustable parameters. This principle of convergence was what Wrinch and I originally called the *simplicity postulate*.† It is not, however, a separate postulate but an immediate application of rule 5. We stated it in a way applicable only to quantitative laws expressed by differential equations, and in *Scientific Inference* I gave a quantitative definition of the complexity of a differential equation. This, however, appears insufficiently general, because it is not clear that all laws are expressible by differential equations: for instance, 'all crows are black', 'the chance of throwing a head with a penny is  $\frac{1}{2}$ ', and the various non-commutative rules of quantum theory. It appears much better not to restrict the possible types of law at all, but merely to be ready for them as they may arise for consideration, whatever their form. This makes the relation to actual thought immediate. The complexity of a law is now merely the number of adjustable parameters in it, and this number is recognizable at once; we can satisfy rule 3. There is no objection to regarding such laws as  $y \propto x$  and  $y \propto x^2$  as of equal complexity, because their consequences will usually differ so much that discrimination between them by means of observations will be easy; laws involving the same number of adjustable parameters can be taken as having the same prior probability. When

† *Phil. Mag.* **42**, 1921, 369-90.

the question of modifying<sup>\*</sup> a law first arises, the suggested modification must be stated, in most cases, in such a form that it involves one new parameter. (A modification to a law of different form, but involving the same number of parameters, can be tested directly. The more probable is the one with the higher likelihood.) The question will then be, Is the new parameter supported by the observations, or is any variation expressible by it better interpreted as random? Thus we must set up two hypotheses for comparison, with equal prior probabilities, so as to say that we have no grounds for expecting it to be present or not.

But if the parameters already introduced are  $\alpha_1, \alpha_2, \dots, \alpha_m$ , and the question is whether we should introduce another,  $\alpha_{m+1}$ , we can choose it so that making it zero will reproduce the old law. This is equivalent, therefore, to saying that we can proceed directly to the law containing  $\alpha_{m+1}$ , but that if we do, half the prior probability is concentrated at  $\alpha_{m+1} = 0$ . We shall see under significance tests how this procedure leads to a test of whether the new parameter is supported by the evidence. At present we need only notice that a parameter that arises in a pure problem of estimation often presupposes a significance test that has disposed of some suggested value that it would have in a simpler law. A significance test itself, if it shows that a new parameter is needed, will lead to an estimate of it on the way. But there are many cases where tests have been applied in analogous cases, or where the evidence is so clear that a quantitative test of significance hardly needs to be applied. For instance, the latitude and longitude of the epicentre and the time of occurrence are obviously relevant parameters to the observations of an earthquake. In a problem of estimation, then, we proceed entirely on the hypothesis that the law is given and that the stated parameters and no others are needed. Their actual values are unknown and our object is to find estimates of them. Though estimation problems really presuppose the solution of the corresponding significance ones, it is convenient to take them first because they are easier mathematically and because in many cases the answer to the significance question is already known.

**3.1.** Our first problem is to find a way of saying that the magnitude of a parameter is unknown, when none of the possible values need special attention. Two rules appear to cover the commonest cases. If the parameter may have any value in a finite range, or from  $-\infty$  to  $+\infty$ , its prior probability should be taken as uniformly distributed. If it

arises in such a way that it may conceivably have any value from 0 to  $\infty$ , the prior probability of its logarithm should be taken as uniformly distributed. There are cases of estimation where a law can be equally well expressed in terms of several different sets of parameters, and it is desirable to have a rule that will lead to the same results whichever set we choose. Otherwise we shall again be in danger of using different rules arbitrarily to suit our taste. It is now known that a rule with this property of invariance exists, and is capable of very wide, though not universal, application.

The essential function of these rules is to provide a formal way of expressing ignorance of the value of the parameter over the range permitted. They make no statement of how frequently that parameter, or other analogous parameters, occur within different ranges. Their function is simply to give formal rules, as impersonal as possible, that will enable the theory to begin. Starting with any distribution of prior probability and taking account of successive batches of data by the principle of inverse probability, we shall in any case be able to develop an account of the corresponding probability at any assigned state of knowledge. There is no logical problem about the intermediate steps that has not already been considered. But there is one at the beginning: how can we assign the prior probability when we know nothing about the value of the parameter, except the very vague knowledge just indicated? The answer is really clear enough when it is recognized that a probability is merely a number associated with a degree of reasonable confidence and has no purpose except to give it a formal expression. If we have no information relevant to the actual value of a parameter, the probability must be chosen so as to express the fact that we have none. It must say nothing about the value of the parameter, except the bare fact that it may possibly, by its very nature, be restricted to lie within certain definite limits.

The uniform distribution of the prior probability was used by Bayes and Laplace in relation to problems of sampling, and by Laplace in some problems of measurement. The problem in sampling would be, given the total number in the population sampled, to use the sample to estimate the numbers of different types in the population. We are prepared for any composition if we know nothing about the population to start with. Hence the rule must be such as to say that we know nothing about it; and Bayes and Laplace did this by taking the prior probabilities of all possible numbers in the population the same and leaving the entire decision to the sample.

Bayes and Laplace, having got so far, unfortunately stopped there, and the weight of their authority seems to have led to the idea that the uniform distribution of the prior probability was a final statement for all problems whatever, and also that it was a necessary part of the principle of inverse probability. There is no more need for the latter idea than there is to say that an oven that has once cooked roast beef can never cook anything but roast beef. The fatal objection to the universal application of the uniform distribution is that it would make any significance test impossible. If a new parameter is being considered, the uniform distribution of prior probability for it would practically always lead to the result that the most probable value is different from zero—the exceptional case being that of a remarkable numerical coincidence. Thus any law expressed in terms of a finite number of parameters would always be rejected when the number of observations comes to be more than the number of parameters determined. In fact, however, the simple rule is retained and the new parameter rejected, at any rate until the latter exceeds a few times its standard error. I maintain that the only ground that we can possibly have for not always rejecting the simple law is that we believe that it is quite likely to be true—that is, that when we have allowed for the variation accounted for by the functions involved in it the rest of the variation is legitimately treated as random, and that we shall get more accurate predictions by proceeding in this way. We do not assert it as certain, but we do seriously consider that it may be true—in other words, it has a non-zero prior probability, which is the prior probability that the new parameter, which is the coefficient of a new function, is zero. But that is a recognition that for the purpose of significance tests, at least, the uniform distribution of the prior probability is invalid.

The uniform distribution of the prior probability was applied to the standard error by Gauss, who, however, seems to have found something unsatisfactory about it. At any rate there is an obvious difficulty. If we take

$$P(d\sigma | H) \propto d\sigma$$

as a statement that  $\sigma$  may have any value between 0 and  $\infty$ , and want to compare probabilities for finite ranges of  $\sigma$ , we must use  $\infty$  instead of 1 to denote certainty on data  $H$ . There is no difficulty in this because the number assigned to certainty is conventional. It is usually convenient to take 1, but there is nothing to say that it always is. But if we take any finite value of  $\sigma$ , say  $\alpha$ , the number for the probability that  $\sigma < \alpha$  will be finite, and the number for  $\sigma > \alpha$  will be infinite. Thus

the rule would say that whatever finite value  $\alpha$  we may choose, if we introduce Convention 3, the probability that  $\sigma < \alpha$  is 0. This is inconsistent with the statement that we know nothing about  $\sigma$ .

This is, I think, the essence of the difficulty about the uniform assessment in problems of estimation. It cannot be applied to a parameter with a semi-infinite range of possible values. Other objections that have been made at various times turn on the point that if a parameter is unknown then any power of it is unknown; but if such a parameter is  $v$ , then if  $v$  lies between  $v_1$  and  $v_1 + dv$ , we should have according to the rule

$$P(v_1 < v < v_1 + dv | H) \propto dv,$$

and if we try to apply the rule also to  $v^n$  we should say also

$$P\{v_1^n < v^n < (v_1 + dv)^n | H\} \propto dv^n \propto v_1^{n-1} dv.$$

The propositions considered on the left are equivalent, but the assessments on the right differ by the variable factor  $v_1^{n-1}$ . There are cases where this problem has arisen. For instance, in the law connecting the mass and volume of a substance it seems equally legitimate to express it in terms of the density or the specific volume, which are reciprocals, and if the uniform rule was adopted for one it would be wrong for the other. Some methods of measuring the charge on an electron give  $e$ , others  $e^2$ ; but  $de$  and  $de^2$  are not proportional. In discussing errors of measurement we do in fact usually represent them in terms of the standard error; but there is no conclusive reason why we should not use the precision constant  $h = 1/\sigma\sqrt{2}$ , and  $d\sigma$  is not proportional to  $dh$ . But while many people had noticed this difficulty about the uniform assessment, they all appear to have thought that it was an essential part of the foundations laid by Laplace that it should be adopted in all cases whatever, regardless of the nature of the problem. The result has been to a very large extent that instead of trying to see whether there was any more satisfactory form of the prior probability, a succession of authors have said that the prior probability is nonsense and therefore that the principle of inverse probability, which cannot work without it, is nonsense too.

The way out is in fact very easy. If  $v\rho$  is constant, then

$$\frac{dv}{v} + \frac{d\rho}{\rho} = 0.$$

If then  $v$  is capable of any value from 0 to  $\infty$ , and we take its prior probability distribution as proportional to  $dv/v$ , then  $\rho$  is also capable

of any value from 0 to  $\infty$ , and if we take its prior probability as proportional to  $d\rho/\rho$  we have two perfectly consistent statements of the same form. Similarly, for any other power,  $dv/v$  and  $dv^n/v^n$  are always proportional, and the constant ratio will be absorbed in the adjustable factor. If we have to express previous ignorance of the value of a quantity over an infinite range, we have seen that to avoid dealing with ratios of infinitesimals we shall have to represent certainty by infinity instead of 1; thus the fact that  $\int_0^\infty dv/v$  diverges at both limits is a satisfactory feature. This argument is equally applicable if  $v$  is restricted to lie between values  $v_1, v_2$ ; for

$$\frac{dv}{v \log(v_2/v_1)} = \frac{dv^n}{v^n \log(v_2^n/v_1^n)}.$$

This point is relevant to the fact that in many practical problems we are not totally ignorant of the standard error when we start. Some knowledge of it is implied by our choice of measuring instruments, which must be capable of reading to less than the standard error and must cover ranges greater than that likely to be covered by the observations. Thus we usually have some vague knowledge initially that fixes upper and lower bounds to the standard error. But  $dv/v$  remains the only rule that is invariant for powers. If in an actual series of observations the standard deviation is much more than the smallest admissible value of  $\sigma$ , and much less than the largest, the truncation of the distribution makes a negligible change in the results.

The point may be put in another way. If a parameter  $v$  is a dimensional magnitude and not a number, and we want to assess  $P(dv | H)$ , where  $H$  contains no information about  $v$  except that it is positive, this can only be of the form  $Av^n dv$ , where  $A$  and  $n$  are constants. For the ratio of two probabilities must be a number, which would not be satisfied if we took the first factor, say, as  $\sin v$ —the sine of a length means nothing. Nor could it be, say,  $e^{-v/a}$ , where  $a$  is some constant of the same dimensions as  $v$ . For then it would assign a definite value to the ratio of the probabilities that  $v$  is less or greater than  $a$ . If, then,  $a$  is known, it contradicts the condition that we know nothing about  $v$  except its existence and that it lies between 0 and  $+\infty$ ; if it is not known we should have to provide a rule for estimating it or for saying that it is unknown, and in either case we are no further forward. The coefficient of  $dv$  must be something that involves no magnitude other than  $v$ , and if  $v$  is dimensional this can be satisfied only by a power of  $v$ . But now if we

consider some fixed value  $a$  the ratio of the probabilities that  $v$  is less or greater than  $a$  is

$$\int_0^a v^n dv \bigg/ \int_a^\infty v^n dv.$$

If  $n > -1$ , the numerator is finite and the denominator infinite. We could therefore introduce Convention 3 and say that the probability that  $v$  is less than any finite value is 0. If  $n < -1$ , the numerator is infinite and the denominator finite, and the rule would say that the probability that  $v$  is greater than any finite value is 0. Both of these would therefore be inconsistent with saying that we know nothing about  $v$ . But if  $n = -1$ , both integrals diverge and the ratio is indeterminate. We cannot now use Convention 3. Thus we attach no value to the probability that  $v$  is greater or less than  $a$ , which is a statement that we know nothing about  $v$  except that it is between 0 and  $\infty$ . Thus the form

$$P(dv | H) \propto dv/v$$

is the only satisfactory one.

I have recently had an objection to it, that if we fix *two* possible values  $a$  and  $b$  the rule will lead to the statement that the probability that  $v$  lies between  $a$  and  $b$  is 0; and it is inferred from this that the rule says that  $v$  is either 0 or  $\infty$  and can have no finite value at all. To the first point I should answer that if we know nothing about  $v$  except that it may have any value over an infinite range we must in any case regard it as a remarkable coincidence if it should be found in a particular arbitrary finite range. If  $a$  and  $b$  are not arbitrary but are suggested by some previous information, then  $v$  is not initially unknown and the previous information should be allowed for. To the second point I should say that what the rule says is that we attach  $\infty$  as the number to represent the total probability of all finite values; it says nothing at all about the probability of an infinite or zero value. It is easy to invent mathematical functions that are everywhere finite but whose integrals diverge, such as

$$f(x) = 1/x \quad (x \neq 0),$$

$$f(x) = 1 \quad (x = 0).$$

Fundamentally the fallacy in the argument is that it assumes the converse of Theorem 2 in the type of case where zero probability does not imply impossibility.

The rule seems to cover all dimensional magnitudes that might conceivably have any value from 0 to  $\infty$ ; and all cases where it appears

equally natural to take a quantity or some power of it as the parameter to be estimated. The extension to all cases where we want to say that a quantity is initially unknown except that it must lie between 0 and  $\infty$  is done by rule 6, that we must introduce the minimum number of independent postulates. If we used a different rule in other such cases we should be making an unnecessary postulate.

If  $P(dv | H) \propto dv/v$ , it is also proportional to  $d \log v$ , and  $\log v$  can have any value from  $-\infty$  to  $+\infty$ . The rule is therefore consistent with the adoption of a uniform distribution for the prior probability of a quantity restricted only to be real. It appears inconsistent at first sight with the uniform assessment for a quantity with a finite range of possible values. If such a quantity is  $x$  and must lie between 0 and 1,  $x/(1-x)$  is a quantity restricted to lie between 0 and  $\infty$ ; which suggests taking a rule suggested by Haldane:

$$P(dx | H) \propto \frac{1-x}{x} d \frac{x}{1-x} \propto \frac{dx}{x(1-x)}.$$

Laplace's and Bayes's assessments in the sampling problem were simply  $dx$ . Haldane's form gives infinite density at the limits. In spite of the apparent inconsistency I think that the  $dv/v$  rule is right; there are better grounds for believing that it says what it is meant to say—that is, nothing—than for the Bayes-Laplace rule. I should not regard the above as showing that  $dx/x(1-x)$  is right for their problem. Other transformations would have the same properties and would be mutually inconsistent if the same rule was taken for all.

I think that at this point we come up against one of the imperfections of the human mind that have given trouble in the theory: that it has an imperfect memory. If everything that attracted its attention was either remembered clearly or completely forgotten it would be much easier to make a formal theory correspond closely to what the mind actually does, and therefore there would be less need for one. Data completely forgotten would then be totally ignored, and we know how to do that; those perfectly remembered could be used in the theory in the usual way. But the mind retains great numbers of vague memories and inferences based on data that have themselves been forgotten, and it is impossible to bring them into a formal theory because they are not sufficiently clearly stated. In practice, if one of them leads to a suggestion of a problem as worth investigating, all that we can do is to treat the matter as if we were approaching it from ignorance—the vague memory is not treated as providing any information at all. If the comment on a competent piece of experimental work, leading to a definite



conclusion, is 'Everybody knew that', the answer is, 'Yes, but nobody knew enough about it to convince anybody else.' Now I am not at all sure that the difficulty about the Bayes-Laplace assessment is not of this kind. Is it a pure statement of ignorance, or has observational evidence, imperfectly catalogued, about the frequency of different sampling ratios in the past somehow got mixed with it? Edgeworth and Pearson held that it was based on the observed fact that sampling ratios had been about uniformly distributed. This might appeal to a meteorologist studying correlations in weather, which do seem to be roughly uniformly distributed over the possible range, but hardly to a Mendelian. Again, is there not a preponderance at the extremes? Certainly if we take the Bayes-Laplace rule right up to the extremes we are led to results that do not correspond to anybody's way of thinking. The rule  $dx/x(1-x)$  goes too far the other way. It would lead to the conclusion that if a sample is of one type with respect to some property there is probability 1 that the whole population is of that type.

It is at least clear that some special hypothesis is needed for quantities that must lie between 0 and 1, for even if we try to obtain a rule by transforming the  $dv/v$  rule the transformation is not unique. A chance or a ratio in a population, if it is treated as unknown, is an adjustable parameter. Now our general considerations showed that an adjustable parameter usually presupposes a significance test that has excluded some suggested value. Is this so here? It appears that it is. Naïve notions of causality would make all population ratios either 0 or 1. On our analysis such a suggestion would never be certain, but we must give it a finite prior probability at the outset. Not to do this goes too far in the opposite direction. Further, though it has been disposed of in many cases, there are, even in our present state of knowledge, many where it appears to be true; apples and oranges do not grow on the same tree. In genetics the suggested values are usually intermediate, such as  $\frac{1}{2}$ ,  $\frac{1}{4}$ , and  $\frac{3}{8}$ ; in such questions as bias of dice they may be  $\frac{1}{6}$  or  $\frac{1}{2}$ . What the suggested values will be in any specific case will depend on the circumstances of the particular problem; we cannot give a universal rule for them beyond the common-sense one, that if anybody does not know what his suggested value is, or whether there is one, he does not know what question he is asking and consequently does not know what his answer means. But then the problem of sampling, as a pure estimation problem, is limited to the case where there is no suggested value and the prior probability has no singularities. Then there is no objection to the uniform distribution, and no other satisfying this condition has

ever been seriously suggested, though there is something to be said for the rule

$$P(dx | H) = \frac{1}{\pi} \frac{dx}{\sqrt{\{x(1-x)\}}}.$$

With this limitation, then, we may as well use the uniform distribution. Even at the present state of knowledge, sampling ratios do seem to be very uniformly distributed except for problems of certain specific types, where suggested values exist. It is not asserted that such a rule will hold for all time, nor can it if the work is done correctly. But we can test what the form suggested would lead to, and say that in the present state of knowledge that is good enough to be going on with.

**3.2. Sampling.** At first I shall extend the Bayes-Laplace theory to the sampling of a finite population. The total number of the population is  $N$ , which will be the sum  $r+s$  of the theory of random sampling. But our problem is now to infer something about  $r$ , given  $N$  and the sampling numbers  $l$  and  $m$ . Hence we must treat  $N$  as given and replace  $s$  by  $N-r$ . Then the probability of the observed numbers, given  $N$  and  $r$ , will be

$$P(l, m | NrH) = {}^rC_l {}^{N-r}C_m / {}^NC_{l+m}. \quad (1)$$

We have no information initially to say that one value of  $r$ , given  $N$ , is more likely than another. Hence we must take all their prior probabilities equal, and

$$P(r | NH) = 1/(N+1). \quad (2)$$

Then by the principle of inverse probability

$$P(r | l, m, N, H) \propto {}^rC_l {}^{N-r}C_m, \quad (3)$$

factors independent of  $r$  having been dropped. But some value of  $r$  in the range 0 to  $N$  inclusive must be the right one, whence

$$\sum_{r=0}^N P(r | l, m, N, H) = 1 \quad (4)$$

and 
$$P(r | l, m, N, H) = {}^rC_l {}^{N-r}C_m / \sum_{r=0}^N {}^rC_l {}^{N-r}C_m. \quad (5)$$

The summation is done by algebraic methods in *Scientific Inference*. A simple alternative way of doing it, suggested to me by Dr. F. J. W. Whipple, is as follows. Suppose that we have a class of  $N+1$  things arranged in a definite order, and that we wish to select  $l+m+1$ . This can be done in  ${}^{N+1}C_{l+m+1}$  ways. But we may proceed as follows. First select an arbitrary member of the class; let it be the  $(r+1)$ th in order. From the remainder we may select  $l$  from those before the  $(r+1)$ th and  $m$  from those after it in  ${}^rC_l {}^{N-r}C_m$  ways. But we might choose any value of  $r$ , and all selections for different values of  $r$  are different,

since the  $(r+1)$ th of the class must be the  $(l+1)$ th of the sample. Hence

$$\sum_{r=0}^N r C_l^{N-r} C_m = {}^{N+1}C_{l+m+1}. \quad (6)$$

If the sample is large and  $N$  is large, the application of Stirling's formula leads to the approximation

$$P(r | l, m, N, H) = \left\{ \frac{n}{2\pi p(1-p)N(N-n)} \right\}^{1/2} \exp \left\{ -\frac{nN\theta^2}{2(N-n)p(1-p)} \right\}, \quad (7)$$

where  $n = l+m$ ;  $p = l/n$ ;  $\theta = \frac{r}{N} - \frac{l}{n}$ . (8)

Thus  $\theta$  measures the departure from proportionality. Its probability is distributed about 0 with standard error  $\{(N-n)p(1-p)/nN\}^{1/2}$ , which approaches  $\{p(1-p)/n\}^{1/2}$  if the population is large compared with the sample. This might be expected from the corresponding result in the direct problem. Further, if  $N/n$  is large the probability of  $l/n$  given  $r/N$  is nearly independent of  $N$ . The sample can therefore give us no information about the size of the population, and the latter is irrelevant to  $r/N$  given the sample, when  $N$  is large. But if  $N$  is such that we must take into account the difference between  $N-n$  and  $N$ , the standard error of  $\theta$  is a little smaller than for a larger population; the solution for the latter would also be applicable to problems of sampling with replacement or of estimating chances. This represents really only the fact that we regard the sample as part of the population, and our definite knowledge of it reduces the standard error of the ratio for a finite population of which the sample is a part.

This may be seen by considering the probability that the *next* specimen will be of the first type. The population being of number  $N$ , of which  $n$  have already been removed, and the members of the first type being  $r$  in number, of which  $l$  have been removed, the probability of the proposition  $p$ , that the next would be of the type, given  $r$ ,  $N$  and the sample, is

$$P(p | l, m, N, r, H) = \frac{r-l}{N-n}. \quad (9)$$

Combining with (5) by the product rule,

$$P(r, p | l, m, N, H) = \frac{r-l}{N-n} r C_l^{N-r} C_m / {}^{N+1}C_{n+1}. \quad (10)$$

The total probability of  $p$  on the data is got by summing over all values of  $r$ . But

$$\frac{r-l}{N-n} \frac{r!}{l!(r-l)!} = \frac{(l+1)r!}{(N-n)(l+1)!(r-l-1)!} = \frac{l+1}{N-n} r C_{l+1} \quad (11)$$

and

$$\sum_{r=0}^N r C_{l+1}^{N-r} C_m^r = {}^{N+1}C_{n+2}. \quad (12)$$

Hence

$$P(p | l, m, N, H) = \frac{l+1}{N-n} \frac{{}^{N+1}C_{n+2}}{{}^{N+1}C_{n+1}} = \frac{l+1}{n+2} = \frac{l+1}{l+m+2}, \quad (13)$$

which is independent of  $N$ . It is usually known as Laplace's rule of succession.<sup>†</sup> Neither Bayes nor Laplace, however, seems to have considered the case of finite  $N$ . They both proceed by considering a chance  $x$ , which would correspond to  $r/N$ , taking the prior probability of  $x$  uniform between 0 and 1, and using the binomial law for the likelihood. The formal result is naturally the same; but I think that the first person to see that the result is independent of  $N$  was Professor C. D. Broad.<sup>‡</sup> Having got so far, we can see at once that the probability, given the sample, that the next  $n'$  will consist of  $l'$  of the first type and  $n'-l'$  of the second is also independent of  $N$ . For we can construct in turn the probabilities of the second further member being of the type, given the sample and the  $(n+1)$ th, of the third given the sample and the  $(n+1)$ th and  $(n+2)$ th, and so on indefinitely. All of these are independent of  $N$ , and the probability of a series of  $l'$  and  $m'$  in any prescribed order will be built up by multiplying the results. This is found to be

$$\frac{(l+1)(l+2)\dots(l+l')(m+1)(m+2)\dots(m+m')}{(l+m+2)(l+m+3)\dots(l+m+l'+m'+1)} \quad (14)$$

irrespective of the order; and the number of possible orders is  ${}^{l'+m'}C_{l'}$ . Hence the probability given the sample that the next  $l'+m'$  will contain just  $l'$  of the first type, in any order, is

$$P(l', m' | l, m, N, H) = \frac{(l'+m')!}{l'! m'!} \frac{(l+1)\dots(l+l')(m+1)\dots(m+m')}{(l+m+2)\dots(l+m+l'+m'+1)}. \quad (15)$$

This leads to some further interesting results. Suppose that  $m = 0$ , so that the sample is all of one type. Then the probability given the sample that the next will be of the type is  $(l+1)/(l+2)$ , which will be large if the sample is large. The probability that the next  $l'$  will all be of the type ( $m' = 0$ ) is  $(l+1)/(l+l'+1)$ . Thus given that all members yet examined are of the type, there is a probability  $\frac{1}{2}$  that the next  $l+1$  will also be of the type; a result given by Pearson by an extension of Laplace's analysis. But if  $l' = N-l$ , the result is  $(l+1)/(N+1)$ . This can be obtained otherwise. For  $l' = N-l$  is the proposition that the entire population is of the same type, and is equivalent to  $r = N$ .

<sup>†</sup> *Mém. de l'Acad. R. d. Sci.*, Paris, 6, 1774, 621; *Œuvres Complètes*, 8, 30. Curiously, it is not reproduced in the *Théorie Analytique*.

<sup>‡</sup> *Mind*, 27, 1918, 389-404.

$$\text{But } P(r = N | l, m, N, H) = {}^N C_l {}^0 C_0 / {}^{N+1} C_{l+1} = \frac{l+1}{N+1}. \quad (16)$$

It follows that with the uniform distribution of the prior probability (1) a large homogeneous sample will establish a high probability that the next member will be of the same type, and a moderate probability that a further sample comparable in size with the first sample will be of the type, (2) sampling will never give a high probability that the whole population is homogeneous unless the sample constitutes a large fraction of the whole population.

3.21. The last result was given by Broad in the paper just mentioned, and was the first clear recognition, I think, of the need to modify the uniform assessment if it was to correspond to actual processes of induction. It was the profound analysis in this paper that led to the work of Wrinch and myself.† We showed that Broad had, if anything, understated his case, and indicated the kind of changes that were needed to meet its requirements. The rule of succession had been generally appealed to as a justification of induction; what Broad showed was that it was no justification whatever for attaching even a moderate probability to a general rule if the possible instances of the rule are many times more numerous than those already investigated. If we are ever to attach a high probability to a general rule, on any practicable amount of evidence, it is necessary that it must have a moderate probability to start with. Thus I may have seen 1 in 1,000 of the 'animals with feathers' in England; on Laplace's theory the probability of the proposition, 'all animals with feathers have beaks', would be about 1/1000. This does not correspond to my state of belief or anybody else's. We might try to avoid the difficulty by introducing testimony, through the principle that if there were animals with feathers and without beaks, somebody would have seen them and I should have heard of it. This is perhaps questionable, but it only shifts the difficulty, because it raises the need to consider the proposition, 'all other people mean the same thing by words as I do', and this would itself be an inductive generalization as hard to accept, on Laplace's theory, as the first. The fundamental trouble is that the prior probabilities  $1/(N+1)$  attached by the theory to the extreme values are so utterly small that they amount to saying, without any evidence at all, that it is practically certain that the population is not homogeneous in respect of the property to be investigated; so nearly certain that no conceivable

† *Phil. Mag.* 42, 1921, 369-90; 45, 1923, 368-74.

amount of observational evidence could appreciably alter this position. The situation is even worse in relation to quantitative laws, as Wrinch and I showed; the extension to continuous magnitudes would make the probability that a new parameter suggested is zero always genuinely infinitesimal, and there would be no way out of the difficulty considered on p. 103. Now I say that for that reason the uniform assessment must be abandoned for ranges including the extreme values, by rule 5 and by the considerations already quoted from Pearson. An adequate theory of scientific investigation must leave it open for any hypothesis whatever that can be clearly stated to be accepted on a moderate amount of evidence. It must not rule out a clearly stated hypothesis, such as that a class is homogeneous, until there is definite evidence against it. Similarly, it must not rule out a quantitative law stated in terms of a finite number of parameters. But this amounts to enunciating the principle: *Any clearly stated law has a finite prior probability, and therefore an appreciable posterior probability until there is definite evidence against it.* This is the fundamental statement of the simplicity postulate. The remarkable thing, indeed, is that this was not seen by Laplace, who in other contexts is referred to as the chief advocate of extreme causality. Had he applied his analysis of sampling to the estimation of the composition of an entire finite population, it seems beyond question that he would have seen that it could never lead to an appreciable probability for a single general law, and is therefore unsatisfactory.

The admission of a probability for the extreme values that remains finite however large the population may be, leads at once to satisfactory results. For if we take

$$P(r = 0 | NH) = P(r = N | NH) = k \quad (17)$$

and distribute the remainder  $1 - 2k$  uniformly over the other values, we shall have

$$P(r | NH) = \frac{1 - 2k}{N - 1} \quad (r \neq 0, N). \quad (18)$$

For  $k = 1/(N + 1)$  this reduces to Laplace's rule. Now if the sample is not homogeneous the extreme possible values of  $N$  give zero probability to the sample, and are therefore excluded by the data; while for comparison of intermediate values the new prior probability merely gives an irrelevant constant factor and leaves the result as it was before. Thus the results derived from a mixed sample will need no change.

But now suppose that the sample is all of the first type, so that  $l = n$ .  $r = 0$  is now excluded by the data, but we want the revised

posterior probability that  $r = N$ . This can be derived easily. For the likelihood factors are unaltered, and for  $r \neq N$  the ratios of the prior probabilities are unaltered. Therefore we need only consider the two alternatives  $r = N$  and  $r \neq 0, N$ , multiplying the previous posterior probabilities in the same ratio as the prior probabilities. The former were  $\frac{n+1}{N+1}$  and  $\frac{N-n}{N+1}$ ; the previous prior probabilities were  $\frac{1}{N+1}$  and  $\frac{N-1}{N+1}$ ; the new prior probabilities  $k$  and  $1-2k$ . Hence, now,

$$\frac{P(r = N | l = n, N, H)}{P(r \neq N | l = n, N, H)} = \frac{n+1}{N-n} \frac{k}{1-2k} \frac{N-1}{1}. \quad (19)$$

Hence if  $n$  is large, the ratio is greater than  $(n+1)k/(1-2k)$  whatever  $N$  may be, and the posterior probability that  $r = N$  will approach 1 as the sample increases, almost irrespective of  $N$ , as soon as  $n$  has reached  $1/k$ . We may notice that if  $n = 1$ , the ratio is  $2k/(1-2k)$ , which is independent of  $N$  if  $k$  is.

The best value to take for  $k$  is not clear, but the following considerations are relevant. If  $k = \frac{1}{2}$ , it says that we already know that  $r = 0$  or  $N$ ; hence this is too large. If  $k = 1/(N+1)$ , we recover the result from the uniform assessment, and this is too low.  $k = \frac{1}{4}$  gives the ratio  $\frac{1}{2}(n+1)\frac{N-1}{N-n}$ , which = 1 if  $n = 1$ ; this would say that a generalization on one instance has probability  $\frac{1}{2}$ , which is not unreasonable. The trouble here is that on the uniform assessment, if  $N = 2$ ,  $k$  is already  $\frac{1}{4}$ , so that  $k = \frac{1}{4}$  is too low in this case. If we are to make a general rule independent of  $N$  we are therefore restricted to values of  $k$  between  $\frac{1}{4}$  and  $\frac{1}{2}$ . A possible alternative form would be to take

$$k = \frac{1}{4} + \frac{1}{2(N+1)}, \quad (20)$$

which puts half the prior probability into the extremes and leaves the other half distributed equally over all values, including the extremes. The basis of such an assessment would be a classification of the possibilities as follows: (1) Population homogeneous on account of some general rule. (2) No general rule, but extreme values to be treated on a level with others. Alternative (1) would then be distributed equally between the two possible cases, and (2) between its  $n+1$  possible cases. This is in accordance with the principles of significance tests, which will be developed later. For  $N = 2$  it gives  $k = \frac{5}{12}$ , leaving  $\frac{1}{6}$  for the prior probability that the two members are unlike. For  $N$  large it

gives the ratio of the posterior probabilities  $\frac{n+1}{2} \frac{N+3}{N-n}$ , which seems satisfactory. It is possible, therefore, to give assessments of the prior probability that avoid the difficulty found by Broad. The solution would be suited to a case where it is still a serious possibility that the class is all of one type, but we do not know of which type.

A partial solution has been given by Pearson.<sup>†</sup> 'Suppose the solidification of hydrogen to have been *once* accomplished. . . . What is the probability that on repetition of the same process the solidification of hydrogen will follow? Now Laplace has asserted that the probability that an event which has occurred  $p$  times and has not hitherto failed will occur again, is represented by the fraction  $\frac{p+1}{p+2}$ . Hence, in the case of hydrogen, the probability of repetition would be only  $\frac{2}{3}$ , or, as we popularly say, the odds would be two to one in its favour. On the other hand, if the sun has risen without fail a million times, the odds in favour of its rising to-morrow would be 1,000,001 to 1. It is clear that on this hypothesis there would be practical certainty with regard to the rising of the sun being repeated, but only some likelihood with regard to the solidification of hydrogen being repeated. The numbers, in fact, do not in the least represent the degrees of belief of the scientist regarding the repetition of the two phenomena. We ought rather to put the problem in this manner:  $p$  different sequences of perceptions have been found to follow the same routine, however often repeated, and none have been known to fail, what is the probability that the  $(p+1)$ th sequence of perceptions will have a routine? Laplace's theorem shows us that the odds are  $p+1$  to 1 in favour of the new sequence having a routine. In other words, since  $p$  represents here the infinite variety of phenomena in which men's past experience has shown that the same causes are on repetition followed by the same effect, there are overwhelming odds that any newly observed phenomenon may be classified under this law of causation. So great and, considering the odds, reasonably great is our belief in this law of causation applying to new phenomena, that when a sequence of perceptions does not appear to repeat itself, we assert with the utmost confidence that the same causes have not been present in the original and in the repeated sequence.' Here Pearson goes far to anticipate the difficulty raised by Broad, in fact too far, for he almost says that exact causality has been established in general by inductive methods. But he has given one

<sup>†</sup> *The Grammar of Science*, 1911, p. 141. Everyman edition, p. 122.



essential point, by transferring the Laplacean inference from simple events to laws. If routines have been established in half the cases already examined, that is adequate ground for attaching a prior probability  $\frac{1}{2}$  that there will be a routine in a new case. If it has been found that all pure substances yet examined have fixed freezing-points, the  $p+1$  to 1 rule would apply as it stands,  $p$  being now the number so far tested. The weakness of the argument is that each of the previous cases of routine has involved an induction from a finite number of observations to a general law, and if we started with the Laplace assessment we should never be able by induction to attach a high probability to even one general law. Pearson's argument, with the above modification, is highly important in relation to present procedure, but the type of assessment (20) is needed at the outset in any case.

3.22. In what follows Dirichlet integrals are used several times. As they are usually expressed in the  $\Gamma$  notation, and I find the factorial notation more convenient (it is also adopted in the British Association Tables), the main formulae are given at this point.

$$\int_0^1 x^l (1-x)^m dx = \frac{l! m!}{(l+m+1)!}. \quad (1)$$

For  $w$  variables all between 0 and 1,

$$\begin{aligned} \int \int \int \dots \int x_1^{l_1} x_2^{l_2} \dots x_w^{l_w} dx_1 \dots dx_w & \quad (0 \leq \sum x \leq 1) \\ &= \frac{l_1! l_2! \dots l_w!}{(l_1 + l_2 + \dots + l_w + w)!}, \end{aligned} \quad (2)$$

$$\begin{aligned} \int \int \int \dots \int x_1^{l_1} x_2^{l_2} \dots x_w^{l_w} dx_1 \dots dx_w & \quad (0 \leq \sum x^p \leq 1) \\ &= \frac{1}{p^w} \frac{\left(\frac{l_1+1}{p}-1\right)! \left(\frac{l_2+1}{p}-1\right)! \dots \left(\frac{l_w+1}{p}-1\right)!}{\left(\frac{l_1+l_2+\dots+l_w+w}{p}\right)!}, \end{aligned} \quad (3)$$

$$\begin{aligned} \int \int \dots \int f(\sum x_i^2) dx_1 \dots dx_w & \quad (0 \leq \sum x_i^2 \leq 1) \\ &= \frac{\pi^{1/2w}}{(\frac{1}{2}w-1)!} \int_0^1 f(u) u^{1/2w-1} du. \end{aligned} \quad (4)$$

For  $l_1 = l_2 = \dots = l_w = 0$ , (2) reduces to  $1/w!$ .

For  $l_1 = l_2 = \dots = l_w = 0$ ,  $p = 2$ , (3) becomes

$$\frac{\{(-\frac{1}{2})!\}^w}{2^w (\frac{1}{2}w)!}.$$

If negative values of the  $x$ 's are admitted, this is multiplied by  $2^w$ . This gives what is often called the volume of a  $w$ -dimensional sphere of radius 1. That of a  $w$ -dimensional sphere of radius  $c$  is therefore

$$\frac{\pi^{1/2w}}{(\frac{1}{2}w)!} c^w, \quad (5)$$

which reduces to  $\pi c^2$  for  $w = 2$ , and  $\frac{4}{3}\pi c^3$  for  $w = 3$ , as it should.

**3.23. Multiple sampling.** When the class sampled consists of several types we can generalize Laplace's assessment, with similar provisos to those needed in the simpler case. Suppose that the whole number of members is  $n$ , divided among  $r$  types, the numbers of the respective types being  $m_1, m_2, \dots, m_r$ . Then we say that all compositions are equally probable. The number of ways of dividing  $n$  things into  $r$  classes is  $(n+r-1)!/n!(r-1)!$ ; but  $m_r$  is determined when the rest are known, and can therefore be omitted by Axiom 6. Hence

$$P(m_1, m_2, \dots, m_{r-1} | nH) = n!(r-1)!/(n+r-1)!. \quad (1)$$

Of these possibilities, if  $m_1$  is considered fixed, the number of partitions among the others is the number of ways of dividing  $n-m_1$  things into  $r-1$  classes, which is  $(n-m_1+r-2)!/(n-m_1)!(r-2)!$ . Hence for  $m_1$  by itself

$$P(m_1 | nH) = \frac{(r-1)n!(n-m_1+r-2)!}{(n+r-1)!(n-m_1)!}. \quad (2)$$

If  $n$  is very large, put  $m_1 = np_1$ , and so on. The proposition that  $m_1$  has a particular value becomes the proposition that  $p_1$  is in a particular range  $dp_1$  of length  $1/n$ . Then

$$\begin{aligned} P(dp_1 dp_2 \dots dp_{r-1} | nH) &= n^{r-1} dp_1 \dots dp_{r-1} \frac{n!(r-1)!}{(n+r-1)!} \\ &\rightarrow (r-1)! dp_1 \dots dp_{r-1}. \end{aligned} \quad (3)$$

Here  $n$  has disappeared and need not be considered further. This gives the distribution of the joint prior probability that the chances of the various types lie in particular ranges. For  $p_1$  separately we can approximate to (2),  $n-m_1$  being large compared with  $r$ , or integrate (3). Then

$$P(dp_1 | H) = (r-1)(1-p_1)^{r-2} dp_1. \quad (4)$$

In (4) the probability of  $p_1$  is no longer uniformly distributed as on Laplace's assessment. This expresses the fact that the average value of all the  $p$ 's is now  $1/r$  instead of  $\frac{1}{2}$  as for the case of two alternatives; it would now be impossible for more than two of them to exceed  $\frac{1}{2}$ . But if all but two of them are fixed the prior probability is uniformly distributed between these two.

Suppose that we have made a sample and that the numbers of various

types are  $x_1, x_2, \dots, x_r$ . The probability of the sample, given the  $p$ 's and the actual order of occurrence, is  $p_1^{x_1} \dots p_r^{x_r}$ ; whence, by (3),

$$P(dp_1 \dots dp_{r-1} | \theta H) \propto p_1^{x_1} \dots p_r^{x_r} dp_1 \dots dp_{r-1}, \quad (5)$$

factors independent of the  $p$ 's having been dropped. Integrating with respect to all  $p$ 's except  $p_1$ , the sum of the others being restricted to be less than  $(1-p_1)$ , we have ( $\theta$  denoting the observed data)

$$\begin{aligned} P(dp_1 | \theta H) &\propto \frac{x_1! \dots x_r!}{(x_1 + \dots + x_r + r - 2)!} p_1^{x_1} (1-p_1)^{x_2 + \dots + x_r + r - 2} dp_1 \\ &\propto p_1^{x_1} (1-p_1)^{x_2 + \dots + x_r + r - 2} dp_1. \end{aligned} \quad (6)$$

But if we are given only  $p_1$ , the probability of getting  $x_1$  of the first type and  $\sum x - x_1$  of the others together is  $p_1^{x_1} (1-p_1)^{\sum x - x_1}$ ; and combining this with (4) we get (6) again, the factor  $r-1$  being independent of  $p_1$ . Hence, if we only want the fraction of the class that is of one particular type, we need consider only the number of that type and the total of the other types in the sample. The distribution among the other types is irrelevant.

By a similar analysis to that used for simple sampling it is found that the probability, given the sample, that the next member chosen will be of the first type is

$$\int_0^1 p_1^{x_1+1} (1-p_1)^{\sum x - x_1 + r - 2} dp_1 \bigg/ \int_0^1 p_1^{x_1} (1-p_1)^{\sum x - x_1 + r - 2} dp_1 = \frac{x_1 + 1}{\sum x + r}. \quad (7)$$

W. E. Johnson,<sup>†</sup> assuming that distribution among the other types is irrelevant to the probability of  $p_1$ , and working entirely with the posterior probability, has shown by an ingenious method that the probability that the next specimen will be of the first type is linear in  $x_1$ . His formula, in the present notation, is  $(wx_1 + 1)/(w \sum x + r)$ .  $w$  is not evaluated; (7) shows that in the conditions considered here  $w = 1$ .

The conditions in question in fact assume that information about the proportion of the class that is of one type is irrelevant to the ratios of the numbers of the other types. They would apply to an estimation of the proportions of blue, white, and pink flowers in *Polygala vulgaris*. We may call this a simple statement of alternatives. If the class falls into main types, according to one set of properties, each of which is subdivided according to another set, and the ratios within one main type give no information about those in another, the result needs some change, as we shall see for a  $2 \times 2$  classification in § 5.11. The numbers of the main types can then be estimated according to Laplace's rule and

<sup>†</sup> *Mind*, 41, 1932, 421-3.

the distribution within each according to that just given. The difference arises from the fact that the discovery that several subtypes of the same main type are rare will give some inductive ground for supposing that other subtypes of that type are also rare: there is no longer complete independence apart from the bare fact that the sum of all the chances must be 1.

**3.3. The Poisson distribution.** The derivation of this law suggests an analogy with sampling, but there is a difference since the one parameter involved is capable of any positive value. It is the product of a chance known to be small in any one trial, and the number of trials, which is large. We might try to regard the problem of radioactivity, for instance, as one of sampling, the problem being to estimate the fraction of the atoms in the specimen that break up in the time of the experiment. But this is not valid because the size of the specimen and the time of the experiment are themselves chosen so as to make the expectation large; we already know that the fraction that break up is small but not zero. This must be expressed by a concentration of the prior probability towards small values. It is not covered by either the uniform assessment or the suggestion of a finite concentration at 0. The fundamental object of the work is to estimate the parameter  $\alpha$  in the formula  $e^{-\alpha t}$ , which represents the fraction of the atoms originally present that survive after time  $t$ . This parameter is not a chance but a chance per unit time, and therefore is dimensional; thus the correct prior probability distribution for it, given that it must lie between 0 and  $\infty$  and is otherwise unknown, is  $d\alpha/\alpha$ . In the dust counter, similarly, the fundamental parameter is the number of particles per unit volume, which again is dimensional; but it might appear equally legitimate to use the mean volume per particle, and the  $dr/r$  rule holds, though possibly with a slight modification to take account of the fact that the air cannot be all dust. In the problem of the soldiers killed by horses a time factor again enters. It appears best, therefore, in problems where the Poisson law arises, to take the prior probability

$$P(dr | H) \propto dr/r. \quad (1)$$

Also given  $r$ , the chance that the event will happen  $m$  times in any interval is

$$P(m | rH) = r^m e^{-r} / m!. \quad (2)$$

The joint chance for several intervals is therefore

$$P(m_1, m_2, \dots, m_n | rH) = \frac{r^{\sum m_i} e^{-nr}}{m_1! m_2! \dots m_n!} \quad (3)$$

and, omitting factors independent of  $r$ , we have

$$P(dr | m_1, m_2, \dots, m_n, H) \propto r^{\sum m - 1} e^{-nr} dr = \frac{n^{\sum m}}{(\sum m - 1)!} r^{\sum m - 1} e^{-nr} dr. \quad (4)$$

The probability, given the observations, that  $r$  is in any particular range is given by the incomplete  $\Gamma$  function.† We notice that the only function of the observations that appears in the posterior probability is  $\sum m$ , which is therefore a sufficient statistic for  $r$ . The utility of further information about the individual  $m$ 's is that they may provide a check on whether the Poisson law actually holds, or whether, for instance, there is a deviation in the direction of the negative binomial. The expectation of  $r$ , given the data, is at  $\bar{m} = (\sum m)/n$ ; the maximum probability density is at a slightly smaller value, and the standard error  $(\bar{m}/n)^{1/2}$  if  $\sum m$  is large.

**3.4. The normal law of error.** We consider first the case where the standard error is known, but the true value  $x$  is unknown over a wide range. Then  $\sigma$  is part of the data  $H$ , and

$$P(dx | H) \propto dx. \quad (1)$$

Also the joint chance of all the observations is

$$P(dx_1 \dots dx_n | x, H) = \frac{1}{(2\pi)^{1/2n} \sigma^n} \exp \left[ -\frac{n}{2\sigma^2} \{(\bar{x} - x)^2 + s'^2\} \right] dx_1 dx_2 \dots dx_n. \quad (2)$$

Hence, omitting factors independent of  $x$ ,‡

$$\begin{aligned} P(dx | x_1, x_2, \dots, x_n, H) &\propto \exp \left\{ -\frac{n}{2\sigma^2} (x - \bar{x})^2 \right\} dx \\ &= \sqrt{\left( \frac{n}{2\pi} \right) \frac{1}{\sigma^2}} \exp \left\{ -\frac{n}{2\sigma^2} (x - \bar{x})^2 \right\} dx, \quad (3) \end{aligned}$$

† J. B. S. Haldane, *Proc. Camb. Phil. Soc.* **28**, 1932, 58. This paper contained the use of the  $dv/v$  rule for the prior probability in such cases, at a time when I had considered it only in relation to a standard error; also the concentration of a finite fraction of the prior probability in a particular value, which later became the basis of my significance tests.

‡ It is understood that  $dx$  in the sign  $P(dx | \dots)$  is an abbreviation for a proposition, namely that a quantity  $\xi$  whose probability distribution is being considered lies in a particular range  $x$  to  $x+dx$ . In the data  $x, H$  of (2),  $x$  is used as an abbreviation for the same proposition; but it is convenient to abbreviate the same proposition in different ways according as it appears in the data or in the proposition whose probability is being considered. The reason is that in (1) or (3)  $P(dx | \dots)$  is an element of a distribution, and the differential calls attention to this fact and appears explicitly on the right; but in (2) the variation of  $x$  in an arbitrarily small range contributes arbitrarily little to the right side, and we need attend only to the value of  $x$ . This method of abbreviation lends itself to immediate adaptation to integration:

$$\int_{x=x_1}^{x_2} P(dx | q) = P(x_1 < x < x_2 | q).$$

so that the posterior probability of  $x$  is normally distributed about  $\bar{x}$  with standard error  $\sigma/\sqrt{n}$ .

In practical cases there is usually some previous information relevant to  $x$ . Perhaps the discovery of a new star (nova) affords the simplest example. The original discovery is a non-quantitative observation, often a naked-eye one, but by comparison with neighbouring stars it gives enough information to enable the observer to identify the new star again. It may be enough to specify the position within  $1^\circ$ , but later measurements may have a standard error of the order of  $1''$ . Then (1) should strictly be replaced by

$$P(dx | H) = f(x) dx,$$

where  $f(x)$  is very small if  $x$  is not within a particular range of order  $1^\circ$ , and within this range  $f(x)$  varies slowly. But then we get

$$P(dx | x_1 \dots x_n, H) \propto f(x) \exp\left\{-\frac{n}{2\sigma^2}(x - \bar{x})^2\right\} dx.$$

$\bar{x}$  is within the range where  $f(x)$  is appreciable (otherwise the accurate observer would be observing the wrong star) and the exponential factor is negligible if  $|x - \bar{x}|$  is more than about  $3''$ . In this range we can neglect the variation of  $f(x)$ , and on adjusting the constant factor we are led again to (3) with a high degree of accuracy. In such cases the original information is not contradicted by the new evidence, but is superseded in the sense that when the latter is available the effect of the original information on the result is negligible. Similar considerations can arise in most of the problems of this chapter and the next, and we shall not usually call special attention to them.

**3.41.** If the standard error is unknown, its prior probability must be proportional to  $d\sigma/\sigma$ , partly because it is usually dimensional and might be either very large or very small, partly because we might equally well take the precision constant as our standard of accuracy. Also we need not suppose that any previous knowledge of  $x$  would tell us anything directly about  $\sigma$ , so that the prior probabilities of  $x$  and  $\sigma$  may be taken independent. Then

$$P(dxd\sigma | H) \propto dxd\sigma/\sigma. \quad (1)$$

The likelihood factor is the same as before; hence

$$P(dxd\sigma | x_1, x_2, \dots, x_n, H) \propto \sigma^{-n-1} \exp\left[-\frac{n}{2\sigma^2}\{(x - \bar{x})^2 + s'^2\}\right] dxd\sigma \quad (2)$$

and the constant factor is

$$\frac{n^{1/2} n_s'^{n-1}}{2^{1/2} n^{-1} \sqrt{\pi} (\frac{1}{2}n - \frac{3}{2})!}.$$

We notice here the immediate representation of the posterior probability in terms of the sufficient statistics  $\bar{x}$  and  $s'$ . All the other factors depending on the observations are the same for all values of  $x$  and  $\sigma$ , and therefore cancel from the posterior probability.

To get the posterior probability of  $x$  by itself we have only to integrate with regard to  $\sigma$ . We have

$$P(dx | x_1, x_2, \dots, x_n, H) \propto dx \int_0^\infty \sigma^{-n-1} \exp \left[ -\frac{n}{2\sigma^2} \{ (x - \bar{x})^2 + s'^2 \} \right] d\sigma \quad (3)$$

which becomes, on putting

$$u = \frac{n}{2\sigma^2} \{ (x - \bar{x})^2 + s'^2 \}, \quad (4)$$

$$P(dx | x_1, x_2, \dots, x_n, H) \propto \left( \frac{2}{n} \right)^{1/2n} \int_0^\infty u^{1/2n} e^{-u} \frac{du}{u} \cdot \{ s'^2 + (x - \bar{x})^2 \}^{-1/2n} dx. \quad (5)$$

Only the last factor involves  $x$ . Determining the constant factor by the condition that  $-\infty < x < \infty$ , we have

$$P(dx | x_1, x_2, \dots, x_n, H) = \frac{1}{\sqrt{\pi}} \frac{(\frac{1}{2}n - 1)!}{(\frac{1}{2}n - \frac{3}{2})!} \frac{s'^{n-1}}{\{ s'^2 + (x - \bar{x})^2 \}^{1/2n}} dx. \quad (6)$$

The right side is identical with 'Student's' rule in form.

Integrating (2) with respect to  $x$  we get

$$P(d\sigma | x_1 \dots x_n, H) \propto \sigma^{-n} \exp \left( -\frac{ns'^2}{2\sigma^2} \right) d\sigma. \quad (7)$$

If  $n = 2$ , and we put  $x - \bar{x} = s' \tan \phi$ , we get

$$P(d\phi | x_1, x_2, H) = \frac{1}{\pi} d\phi. \quad (8)$$

But in this case  $s'$  is simply the distance of either observation from the mean, and the values  $\phi = \pm \frac{1}{4}\pi$  give, respectively,  $x = x_1$  and  $x = x_2$ . Hence

$$P(x_1 < x < x_2 | x_1, x_2, H) = \frac{1}{2}. \quad (9)$$

That is, given just two observations the true value is as likely as not to lie between them. This is a general result for any type of law that involves only a location and a scale parameter, both of which are initially unknown. The latter condition is necessary. If, for instance,  $H$  contained information about the standard error, and the first two observations differed by  $4\sigma$ , there would be a high probability, given these observations, that the true value was about midway between them, and then the probability that the true value was between them

would be more than  $\frac{1}{2}$ . If they differed by  $\frac{1}{2}\sigma$ , on the other hand, we should interpret this as an accidental agreement and the probability, given the observations, that the true value lay between them would be less than  $\frac{1}{2}$ . It is only when the observations contain the whole of the information available about  $\sigma$  that the probability, given them, that the true value lies between them can be the same for all possible separations of the observations.

If  $n = 1$ ,  $\bar{x} = x_1$ , and  $s' = 0$ . Then returning to (2)

$$P(dx d\sigma | x_1, H) \propto \sigma^{-2} \exp\left\{-\frac{(x-\bar{x})^2}{\sigma^2}\right\} dx d\sigma. \quad (10)$$

Integrating with regard to  $\sigma$  we get

$$P(dx | x_1, H) \propto \frac{dx}{|x-x_1|}, \quad (11)$$

that is, the most probable value of  $x$  is  $x_1$ , but we have no information about the accuracy of the determination. (7) gives for  $\sigma$

$$P(d\sigma | x_1, H) \propto d\sigma/\sigma, \quad (12)$$

that is, we still know nothing about  $\sigma$ . These results were to be expected, but attention to degenerate cases is often desirable to verify that the solution does degenerate in the right way.

It is easy to show that, with the distribution of probability given in (6), the expectation of  $(x-\bar{x})^2$  is, for  $n > 3$ ,

$$\frac{s'^2}{n-3} = \frac{\sum (x_r - \bar{x})^2}{n(n-3)}, \quad (13)$$

and is infinite if  $n$  is less than 4. At small numbers of observations the departure of the posterior probability from normality is great.

There is, however, the following peculiarity if two sets of observations with different standard errors  $\sigma$ ,  $\tau$  are relevant to the same  $x$ . We should here take

$$P(dx d\sigma d\tau | H) \propto dx d\sigma d\tau / \sigma \tau,$$

$$P(\theta | x, \sigma, \tau, H) \propto \sigma^{-m} \tau^{-n} \exp\left[-\frac{m}{2\sigma^2}\{s'^2 + (x-\bar{x})^2\} - \frac{n}{2\tau^2}\{t'^2 + (x-\bar{y})^2\}\right].$$

Combining these and integrating with regard to  $\sigma$  and  $\tau$ , we get

$$P(dx | \theta H) \propto \{s'^2 + (x-\bar{x})^2\}^{-1/2m} \{t'^2 + (x-\bar{y})^2\}^{-1/2n} dx, \quad (14)$$

and the expectation of  $x^2$  converges even if  $m = n = 2$ . The integral needs complicated elliptic functions to express it if  $m$  and  $n$  are odd, and in general is not expressible compactly. If  $m = 1$ ,  $n = 2$  we find that the posterior probability has a pole at  $\bar{x}$ , but the expectation of



$(x - \bar{x})^2$  is infinite; this means that neither very small nor very large values of  $\sigma$  are yet effectively excluded by the data.

If an estimate has standard error  $\sigma$ ,  $\sigma^{-2}$  or some number proportional to it is called the *weight*. If  $x_1, x_2, \dots$  are a set of estimates of  $x$ , with weights  $w_1, w_2, \dots$ , the most probable value of  $x$  is given by

$$x \sum w_r = \sum w_r x_r, \quad (15)$$

and if unit weight corresponds to standard error 1, the standard error of the estimate is  $(\sum w_r)^{-1/2}$ . This additive property of weight often makes it convenient to express the standard errors in terms of it. The standard error, itself, however, has an additive property. If  $x_1$  and  $x_2$  have independent standard errors  $\sigma_1$  and  $\sigma_2$ , then the standard error of  $x_1 + x_2$  or of  $x_1 - x_2$  is  $(\sigma_1^2 + \sigma_2^2)^{1/2}$ , and the corresponding weight is  $w_1 w_2 / (w_1 + w_2)$ .

The usual practice in astronomical and physical work is to multiply the estimated standard error by 0.6745 and call the result the 'probable error'. But this multiplication, which has little point even when the probability considered is normally distributed, is seriously wrong when uncertainty is estimated from the observations themselves. Writing the usual estimate of the standard error in the form

$$s_x = \left\{ \sum \frac{(x_r - \bar{x})^2}{n(n-1)} \right\}^{1/2} \quad (16)$$

$$\text{and} \quad t = (x - \bar{x}) / s_x, \quad (17)$$

we find as for 2.8(21)

$$P(dt | \theta H) \propto \left( 1 + \frac{t^2}{n-1} \right)^{-1/2n} dt, \quad (18)$$

which is not normal. We have already seen that for  $n = 2$  the probability that  $x$  is between  $\bar{x} \pm s_x$  is  $\frac{1}{2}$ , so that the probable error in the sense defined for the normal law is equal to the standard error. For risks of larger error the difference is greater.  $P$  being the probability of a larger  $t$  (positive and negative errors being taken together) we have the following specimen values, from Fisher's table.

$n \backslash P$	0.5	0.1	0.05	0.01
2	1.000	6.314	12.706	63.657
5	0.727	2.132	2.776	4.604
10	0.703	1.833	2.262	3.250
20	0.688	1.729	2.093	2.861
$\infty$	0.674	1.645	1.960	2.576

The values depart widely from proportionality, and a statement of uncertainty based on only a few observations is useless as a guide to the risk of large error unless the number of observations is given.

In many statements of the results of physical experiments, besides the omission of explicit statement of the numbers of observations in the final conclusion, the uncertainties stated are often rounded to one figure; I have actually seen a 'probable error' given as 0.1, which might mean anything from 0.05 to 0.15. Suppose then that two estimated standard errors are both given as 0.1, but one means 0.05 on 20 observations, the other 0.15 on 2 observations; and that we want to state limits such that there is a probability 0.99 that the true value lies between them—which we might quite well want to do if much depends on the answer. The limit in one case would be 0.14, in the other 9.5. In fact if anybody wants to reduce a good set of observations to meaninglessness he can hardly do better than to round the uncertainty to one figure and suppress the number of observations.

It is generally enough to give two figures in the estimated standard error. Karl Pearson usually gave many more figures, often six or seven, and statisticians still usually give four, but I consider more than two a waste of labour. It is not often that a result of importance depends on whether the standard error is 0.95 or 1.05.

**3.42.** The following problem is liable to arise in practice. Given one set of observations derived from the normal law, say  $x_1$  to  $x_{n_1}$ , and no other information about  $x$  and  $\sigma$ , what is the probability that a new series of  $n_2$  observations will give a mean or a standard deviation in a particular range? We have, from 2.8 (15),

$$P(d\bar{x}_2 ds'_2 | x, \sigma, H) = \sqrt{\left(\frac{n_2}{2\pi}\right) \frac{1}{\sigma}} \exp\left\{-\frac{n_2}{2\sigma^2}(\bar{x}_2 - x)^2\right\} d\bar{x}_2 \frac{n_2^{1/2} n_1 - 1/2 s'_2 n_1 - 2}{2^{1/2} n_1 - 3/2 (\frac{1}{2} n_2 - \frac{3}{2})! \sigma^{n_1-1}} \exp\left(-\frac{n_2 s'^2_2}{2\sigma^2}\right) ds'_2 \quad (1)$$

and from 3.41 (2)

$$P(dxd\sigma | x_1, \dots, x_{n_1}, H) = \frac{n_1^{1/2} n_1}{2^{1/2} n_1 - 1 \sqrt{\pi} (\frac{1}{2} n_1 - \frac{3}{2})!} \frac{s'^{n_1-1}_1}{\sigma^{n_1+1}} \exp\left[-\frac{n_1}{2\sigma^2}\{(x - \bar{x}_1)^2 + s'^2_1\}\right] dxd\sigma, \quad (2)$$

whence

$$P(dxd\sigma d\bar{x}_2 ds'_2 | x_1, \dots, x_{n_1}, H) = \frac{n_1^{1/2} n_1 n_2^{1/2} n_1 s'^{n_1-1}_1 s'^{n_2-2}_2}{2^{1/2} n_1 + 1/2 n_2 - 2 \pi (\frac{1}{2} n_1 - \frac{3}{2})! (\frac{1}{2} n_2 - \frac{3}{2})! \sigma^{n_1+n_2+1}} \times \\ \times \exp\left[-\frac{n_1}{2\sigma^2}\{(x - \bar{x}_1)^2 + s'^2_1\} - \frac{n_2}{2\sigma^2}\{(x - \bar{x}_2)^2 + s'^2_2\}\right] dxd\sigma d\bar{x}_2 ds'_2. \quad (3)$$

But

$$n_1(x - \bar{x}_1)^2 + n_2(x - \bar{x}_2)^2 = (n_1 + n_2) \left(x - \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}\right)^2 + \frac{n_1 n_2}{n_1 + n_2} (\bar{x}_1 - \bar{x}_2)^2 \quad (4)$$

and integration with regard to  $x$  gives

$$P(d\sigma d\bar{x}_2 ds'_2 | x_1, \dots, x_n, H) \\ = \frac{n_1^{1/2} n_2^{1/2} s_1'^{n_1-1} s_2'^{n_2-2}}{2^{1/2} n_1 + 1/2 n_2 - 3/2 \sqrt{\pi} (\frac{1}{2} n_1 - \frac{3}{2})! (\frac{1}{2} n_2 - \frac{3}{2})! \sigma^{n_1+n_2} (n_1+n_2)^{1/2}} \times \\ \times \exp\left\{-\frac{n_1 n_2}{2(n_1+n_2)\sigma^2} (\bar{x}_2 - \bar{x}_1)^2\right\} \exp\left\{-\frac{1}{2\sigma^2} (n_1 s_1'^2 + n_2 s_2'^2)\right\} d\sigma d\bar{x}_2 ds'_2. \quad (5)$$

If we now integrate with regard to  $\sigma$ , a factor

$$\left\{n_1 s_1'^2 + n_2 s_2'^2 + \frac{n_1 n_2}{n_1+n_2} (\bar{x}_2 - \bar{x}_1)^2\right\}^{-1/2(n_1+n_2-1)} \quad (6)$$

will arise. This does not separate into factors. Hence, given  $\bar{x}_1$  and  $s'_1$ , the probability distributions of  $\bar{x}_2$  and  $s'_2$  are not independent; though they are independent given  $x$  and  $\sigma$ . What this means is that if  $s'_1$  is unusually large in comparison with  $\sigma$ , we shall overestimate the scale, and this overestimation will affect the estimates of the ranges likely both for  $\bar{x}_2$  and  $s'_2$ . But if we are interested in only one of them we can integrate with regard to the other. Then

$$P(d\sigma d\bar{x}_2 | x_1, \dots, x_n, H) \\ = \frac{n_1^{1/2} n_2^{1/2} s_1'^{n_1-1}}{2^{1/2} n_1 - 1/2 \sqrt{\pi} (\frac{1}{2} n_1 - \frac{3}{2})! (n_1+n_2)^{1/2} \sigma^{n_1+1}} \exp\left\{-\frac{n_1 n_2 (\bar{x}_2 - \bar{x}_1)^2}{2(n_1+n_2)\sigma^2} - \frac{n_1 s_1'^2}{2\sigma^2}\right\} d\sigma d\bar{x}_2, \quad (7)$$

$$P(d\bar{x}_2 | x_1, \dots, x_n, H) = \frac{n_2^{1/2}}{(n_1+n_2)^{1/2}} \frac{(\frac{1}{2} n_1 - 1)!}{\sqrt{\pi} (\frac{1}{2} n_1 - \frac{3}{2})!} \left\{1 + \frac{n_2 (\bar{x}_2 - \bar{x}_1)^2}{(n_1+n_2) s_1'^2}\right\}^{-1/2 n_1} \frac{d\bar{x}_2}{s_1'}. \quad (8)$$

Also

$$P(d\sigma ds'_2 | x_1, \dots, x_n, H) \\ = \frac{n_1^{1/2} n_2^{1/2} s_1'^{n_1-1} s_2'^{n_2-2}}{2^{1/2} n_1 + 1/2 n_2 - 3/2 (\frac{1}{2} n_1 - \frac{3}{2})! (\frac{1}{2} n_2 - \frac{3}{2})! \sigma^{n_1+n_2-1}} \exp\left\{-\frac{n_1 s_1'^2 + n_2 s_2'^2}{2\sigma^2}\right\} d\sigma ds'_2, \quad (9)$$

$$P(ds'_2 | x_1, \dots, x_n, H) = \frac{2 n_1^{1/2} n_2^{1/2} s_1'^{n_1-1} s_2'^{n_2-2} (\frac{1}{2} n_1 + \frac{1}{2} n_2 - 2)! s_1'^{n_1-1} s_2'^{n_2-2}}{(\frac{1}{2} n_1 - \frac{3}{2})! (\frac{1}{2} n_2 - \frac{3}{2})! (n_1 s_1'^2 + n_2 s_2'^2)^{1/2 n_1 + 1/2 n_2 - 1}} ds'_2, \quad (10)$$

$$\text{and on putting} \quad s'_2 = s'_1/y, \quad (11)$$

we recover the form 2.81 (24), and the  $z$  distribution follows.

**3.43.** If  $\bar{x}_2, \dots, \bar{x}_{r+1}$  are the means of  $r$  further sets of  $n_2$  observations each,  $s'_2, \dots, s'_{r+1}$  the corresponding mean square deviations, the rule holds for each separately and independently. Hence

$$P(d\bar{x}_2 \dots d\bar{x}_{r+1} | x, \sigma, H) = \left(\frac{n_2}{2\pi}\right)^{1/2 r} \frac{1}{\sigma^r} \exp\left\{-\frac{n_2}{2\sigma^2} \sum (\bar{x}_m - x)^2\right\} d\bar{x}_2 \dots d\bar{x}_{r+1}. \quad (12)$$

Now put 
$$\sum \bar{x}_m = \bar{r}X, \quad \sum (\bar{x}_m - \bar{x})^2 = rS^2; \quad (13)$$

the exponent becomes 
$$-\frac{rn_2}{2\sigma^2}\{(X-x)^2 + S^2\} \quad (14)$$

and (12) is of exactly the form of 2.8(4), with  $r$  written for  $n$  and  $\sigma/\sqrt{n_2}$  for  $\sigma$ . Hence (10) and the  $z$  rule are adapted immediately to give the probability distribution of  $S$  given  $x_1$  to  $x_{n_1}$ . We need only replace  $n_2$  by  $r$  and  $s_2'^2$  by  $n_2 S^2$ .

This form is more closely analogous to the way in which the  $z$  rule is used in agricultural experiments. In them the means of plots with the same treatment are taken, and the sum of the squares of the differences between the treatment means and the general mean gives  $rS^2$ ;  $n_2 rS^2$  is called the treatment sum of squares. The differences not explicable by treatments or other systematic effects are used to provide  $s_1'$ . Then, given  $s_1'$  and the hypothesis of general randomness, the method will give the probability distribution of  $S$ . If the observed value is such that it would be very unlikely to occur on this hypothesis, given  $s_1'$ , then the hypothesis is rejected and the existence of treatment differences asserted. In Fisher's form  $s_2$  would correspond to the random variation and  $s_1$  to the possibly partly or mainly systematic one, hence his convention that  $s_1 > s_2$ . It is easy to see that interchanging  $n_1$  and  $s_1$  with  $n_2$  and  $s_2$ , and reversing the sign of  $z$ , leaves 2.81 (26) unaltered.

These results were obtained by Mr. W. O. Storer in an unpublished paper, based on a suggestion of mine that the conditions that lead to the similarity between 'Student's' result and mine seemed to be fulfilled also in the circumstances considered by Fisher in deriving the  $z$  distribution. Hence I expected that the probability distribution of  $\log(s_2/s_1)$ , given one set of observations, would agree exactly with that derived from Fisher's formula; and Storer found this to be the case.

**3.44.** A closely related problem, which will serve as an introduction to some features of the method of least squares, is where we have to estimate  $m$  unknowns  $x_r$  ( $r = 1$  to  $m$ ), to each of which a set of  $n_r$  measures  $x_{ri}$  ( $i = 1$  to  $n_r$ ) is relevant. The standard error of one observation is supposed to be the same in all series. Put,  $S$  denoting summation with regard to  $i$ ,  $\Sigma$  with regard to  $r$ ,

$$n_r \bar{x}_r = Sx_{ri}, \quad n_r s_r'^2 = S(x_{ri} - \bar{x}_r)^2. \quad (1)$$

Then, denoting the observations collectively by  $\theta$ ,

$$P(dx_1 \dots dx_m d\sigma | H) \propto dx_1 \dots dx_m d\sigma / \sigma, \quad (2)$$

$$P(\theta | x_1 \dots x_m \sigma H) \propto \sigma^{-\Sigma n_r} \prod_r \exp \left[ -\frac{n_r}{2\sigma^2} \{ (x_r - \bar{x}_r)^2 + s_r'^2 \} \right], \quad (3)$$

$$P(dx_1 \dots dx_m d\sigma | \theta H) \propto \sigma^{-\Sigma n_r - 1} \exp \left[ -\sum \frac{n_r}{2\sigma^2} \{ (x_r - \bar{x}_r)^2 + s_r'^2 \} \right] \prod dx_r d\sigma. \quad (4)$$

By integration

$$P(d\sigma | \theta H) \propto \sigma^{-\Sigma n_r + m - 1} \exp \left\{ -\sum \left( \frac{n_r s_r'^2}{2\sigma^2} \right) \right\} d\sigma, \quad (5)$$

and if we now put

$$\sum (n_r s_r'^2) = (\sum n_r - m) s^2 \quad (6)$$

$$P(d\sigma | \theta H) \propto \sigma^{-(\Sigma n_r - m + 1)} \exp \left\{ -\frac{(\sum n_r - m) s^2}{2\sigma^2} \right\} d\sigma. \quad (7)$$

This is of the same form as 3.41 (7), if in the latter we replace  $ns'^2$  by  $(n-1)s^2$  and then replace  $n-1$  by  $\sum n_r - m$ . In the former problem  $n-1$ , in the present one  $\sum n_r - m$ , is the difference between the whole number of observations and the number of true values to be estimated. Hence it is convenient to call this difference the *number of degrees of freedom* and to denote it by  $\nu$ , and to give the name *standard deviation* to  $s$  in both cases. Then however many unknowns are estimated we always have

$$P(d\sigma | \theta H) \propto \sigma^{-(\nu+1)} \exp \left( -\frac{\nu s^2}{2\sigma^2} \right) d\sigma, \quad (8)$$

and the posterior probability distribution of  $\sigma/s$  is given by a single set of tables.

(4) can now be written

$$P(dx_1 \dots dx_m d\sigma | \theta H) \propto \sigma^{-\Sigma n_r - 1} \exp \left\{ -\sum \frac{n_r}{2\sigma^2} (x_r - \bar{x}_r)^2 - \frac{\nu s^2}{2\sigma^2} \right\} \prod dx_r d\sigma. \quad (9)$$

Integrate with respect to  $x_2, \dots, x_m$ ; then

$$P(dx_1 d\sigma | \theta H) \propto \sigma^{-\Sigma n_r + m - 2} \exp \left\{ -\frac{n_1}{2\sigma^2} (x_1 - \bar{x}_1)^2 - \frac{\nu s^2}{2\sigma^2} \right\} dx_1 d\sigma, \quad (10)$$

$$P(dx_1 | \theta H) \propto \{ \nu s^2 + n_1 (x_1 - \bar{x}_1)^2 \}^{-1/2(\nu+1)} dx_1. \quad (11)$$

Put

$$s_{x_1} = s/\sqrt{n_1}, \quad (12)$$

then

$$P(dx_1 | \theta H) \propto \left\{ 1 + \frac{(x_1 - \bar{x}_1)^2}{\nu s_{x_1}^2} \right\}^{-1/2(\nu+1)} dx_1. \quad (13)$$

Hence the posterior probability of  $x_1$  follows the  $t$  rule with  $\nu$  degrees of freedom, where

$$t = (x_1 - \bar{x}_1)/s_{x_1}. \quad (14)$$

$s_{x_1}$  is related to  $s$  in the same way as the standard error of  $\bar{x}_1$  would be to that of one observation if the latter was known accurately. Hence it is convenient to quote  $s_{x_1}$  as the standard error of  $x_1$ ;  $\bar{x}_1$ ,  $s_{x_1}$ , and  $\nu$

are enough to specify the posterior probability of  $x_1$  completely, while  $s$  and  $\nu$  give that of  $\sigma$  completely.

The situation considered is a common one in practice. A large number of unknowns may have to be estimated, but the number of observations directly relating to any one may be small. The estimate of any unknown from the observations directly relating to it may be of very doubtful accuracy on account of the small number of degrees of freedom. But if the standard error may be assumed the same for observations of all sets the number of degrees of freedom is much increased and a good determination of accuracy becomes possible.

As an example we take Bullard's observations of gravity† in East Africa. Seven stations were visited twice or more, many others only once. Those visited more than once were as follows:

	$g$ (cm./sec. <sup>2</sup> )	Mean	Residual (10 <sup>-4</sup> cm./sec. <sup>2</sup> )
Nakuru . .	977.4810 .4800	977.4805	+5 -5
Kisumu . .	977.6056 .6045	977.6050	+6 -5
Equator . .	977.2608 .2602	977.2605	+3 -3
Mombasa. .	977.0212 .0242	977.0227	-15 +15
Jinja . .	977.7186 .7176	977.7182	+4 -6
Nairobi . .	977.5289 .5307	977.5292	+1 -3
Naivasha. .	977.4663 .4695	977.4679	-11 -16 +16

The sum of squares of residuals is 1499;  $\nu = 16 - 7 = 9$ ; hence

$$10^4 s = (1499/9)^{1/2} = 12.9.$$

Then

$$\begin{aligned} s_{x_r} &= 0.00129 \left( 1, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{3}} \right) \text{ cm./sec.}^2 \\ &= (0.0013, 0.00091, 0.00074) \text{ cm./sec.}^2 \end{aligned}$$

according as the number of measures at a station was 1, 2, or 3; in each case based on 9 d.f.

**3.5. The method of least squares.** This is the extension of the problem of estimation, given the normal law of error, to the case where

† *Phil. Trans. A*, 235, 1936, 445-531.

several unknowns besides, usually, the standard error need to be found. If the unknowns are  $x_i$ ,  $m$  in number, and a measure is  $c_r$ , then if there were no random error we should have a set of relations of the form

$$c_r = f_r(x_1, x_2, \dots, x_m). \quad (1)$$

Actually, on account of the random error, this must be replaced by

$$P(dc_r | x_i, \sigma, H) = \frac{1}{\sqrt{(2\pi)\sigma}} \exp\left\{-\frac{1}{2\sigma^2}(c_r - f_r)^2\right\} dc_r, \quad (2)$$

and if there are  $n$  observations whose errors are independent we can denote them collectively by  $\theta$  and write

$$P(\theta | x_i, \sigma, H) = \frac{1}{(2\pi)^{1/2n}\sigma^n} \exp\left\{-\frac{1}{2\sigma^2}S(c_r - f_r)^2\right\} dc_1 \dots dc_n, \quad (3)$$

$S$  denoting summation over the observations. Usually the functions  $f_r$  are either linear, or else we can find an approximate set of values of the  $x_i$ , say  $x_{i0}$ , and treat the actual  $x_i$  as the sum of  $x_{i0}$  and a small departure  $x'_i$ . In the latter case we can take  $x'_i$  as a new set of unknowns, so that within the permitted range  $\partial f_r / \partial x'_i$  can be treated as constants. In either case we can write

$$W = \frac{1}{2}S(c_r - f_r)^2, \quad (4)$$

which will be a quadratic function of  $x_i$  or of  $x'_i$ . The accent can now be omitted. We can also write

$$f_r = \sum a_{ir} x_i, \quad (5)$$

$\Sigma$  denoting summation over the unknowns; but we can shorten the writing by using the summation convention that when a suffix  $i$  is repeated it is to be given all values from 1 to  $m$  and the results added. To avoid confusion through a suffix occurring more than twice we now write

$$W = \frac{1}{2}S(a_{ir}x_i - c_r)(a_{jr}x_j - c_r) \quad (6)$$

$$= \frac{1}{2}S(a_{ir}a_{jr}x_i x_j - 2a_{ir}c_r x_i + c_r^2) \quad (7)$$

$$= \frac{1}{2}b_{ij}x_i x_j - d_i x_i + \frac{1}{2}Sc_r^2. \quad (8)$$

In the first sum each pair of unequal suffixes occurs twice, since either may be called  $i$  and the other  $j$ . There is always a set of values of  $x$ ; that make  $W$  a minimum. If we call these  $y_i$ , differentiate with regard to  $x_i$ , and put  $y_j$  for  $x_j$ , we have  $m$  equations

$$b_{ij}y_j - d_i = 0. \quad (9)$$

These are called the normal equations. They have a unique solution if  $m \leq n$  and the determinant formed by the  $b_{ij}$  is not zero. Put

$$x_i = y_i + z_i; \quad c_r - a_{ir}y_i = c'_r. \quad (10)$$

Then  $W$  is quadratic in  $z_i$ , and its first derivatives with regard to  $z_i$  all vanish when the  $z_i$  are 0. Also  $W$  is then equal to  $\frac{1}{2}Sc_r'^2$ . Hence

$$W = \frac{1}{2}b_{ij}z_i z_j + \frac{1}{2}Sc_r'^2. \quad (11)$$

Also  $b_{ij}z_i z_j$  is essentially positive because it is equal to  $S(a_{ir}z_i)^2$ ; and it can be reduced to the sum of  $m$  squares of linear functions in an infinite number of ways. The most convenient is illustrated most easily by the case of three unknowns. Suppose

$$F = b_{11}z_1^2 + 2b_{12}z_1 z_2 + b_{22}z_2^2 + 2b_{13}z_1 z_3 + 2b_{23}z_2 z_3 + b_{33}z_3^2. \quad (12)$$

$$\text{Take} \quad \zeta_1 = z_1 + \frac{b_{12}}{b_{11}}z_2 + \frac{b_{13}}{b_{11}}z_3. \quad (13)$$

Then

$$\begin{aligned} F - b_{11}\zeta_1^2 &= \left(b_{22} - \frac{b_{12}^2}{b_{11}}\right)z_2^2 + 2\left(b_{23} - \frac{b_{12}b_{13}}{b_{11}}\right)z_2 z_3 + \left(b_{33} - \frac{b_{13}^2}{b_{11}}\right)z_3^2 \\ &= b'_{22}z_2^2 + 2b'_{23}z_2 z_3 + b'_{33}z_3^2. \end{aligned} \quad (14)$$

Now put

$$\zeta_2 = z_2 + \frac{b'_{23}}{b'_{22}}z_3, \quad (15)$$

$$F - b_{11}\zeta_1^2 - b'_{22}\zeta_2^2 = \left(b'_{33} - \frac{b_{23}^2}{b'_{22}}\right)z_3^2 = b''_{33}z_3^2. \quad (16)$$

The process can evidently be extended to any number of unknowns.

First suppose that  $\sigma$  is known, and take the prior probabilities of  $x_1, \dots, x_m$  uniformly distributed. Then

$$P(dx_1 dx_2 \dots dx_m | \sigma, H) \propto dx_1 \dots dx_m. \quad (17)$$

$$\begin{aligned} P(dx_1 \dots dx_m | \theta, \sigma, H) &\propto \sigma^{-n} \exp\left(-\frac{W}{\sigma^2}\right) dx_1 \dots dx_m \\ &\propto \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}(b_{ij}z_i z_j + Sc_r'^2)\right\} dx_1 \dots dx_m. \end{aligned} \quad (18)$$

But by the mode of formation of the  $\zeta_i$  we see that in the Jacobian  $\frac{\partial(\zeta_1, \dots, \zeta_m)}{\partial(z_1, \dots, z_m)}$  all terms in the leading diagonal are 1, and all those to one side of it are 0. Hence the Jacobian is 1, and we have the form

$$P(dx_1 \dots dx_m | \theta, \sigma, H) \propto \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}(\sum b_i \zeta_i^2 + Sc_r'^2)\right\} d\zeta_1 \dots d\zeta_m. \quad (19)$$

This breaks up into factors, and we can say for any  $\zeta_i$  separately

$$P(d\zeta_i | \theta, \sigma, H) \propto \frac{1}{\sqrt{(2\pi/b_i)\sigma}} \exp\left(-\frac{b_i \zeta_i^2}{2\sigma^2}\right) d\zeta_i. \quad (20)$$



In particular, since  $\zeta_m = z_m$ , we shall be able to write

$$x_m = y_m + z_m = y_m \pm \sigma/\sqrt{b_m}. \quad (21)$$

$b_m$  can be identified easily, for if we write

$$D = ||b_{ij}|| \quad (22)$$

for the determinant of all the  $b_{ij}$ , and  $B_{mm}$  for the minor of  $b_{mm}$  in it, the transformation alters neither  $D$  nor  $B_{mm}$ , since

$$\frac{\partial(\zeta_1, \dots, \zeta_m)}{\partial(z_1, \dots, z_m)} = 1, \quad \frac{\partial(\zeta_1, \dots, \zeta_{m-1})}{\partial(z_1, \dots, z_{m-1})} = 1, \quad (23)$$

and therefore

$$b_m = D/B_{mm}. \quad (24)$$

Any other function of the  $x_i$  can be estimated as follows. Let

$$\xi = l_i x_i = l_i y_i + l_i z_i, \quad (25)$$

where the  $l_i$  are specified. Then we can eliminate the  $z_i$  in favour of the  $\zeta_i$ , and get

$$\xi = l_i y_i + \lambda_i \zeta_i, \quad (26)$$

where the probability of  $\zeta_i$  is distributed about 0 with standard error  $\sigma/\sqrt{b_i}$ . Hence that of  $\xi$  is distributed about  $l_i y_i$  with standard error  $\sigma(\xi)$  given by

$$\sigma^2(\xi) = \sigma^2 \sum (\lambda_i^2/b_i). \quad (27)$$

If  $\sigma$  is unknown we must replace (17) by

$$P(dx_1 \dots dx_m d\sigma | H) \propto dx_1 \dots dx_m d\sigma / \sigma \quad (28)$$

and (19) by

$$P(dx_1 \dots dx_m d\sigma | \theta H) \propto \sigma^{-n-1} \exp\left\{-\frac{1}{2\sigma^2}(\sum b_i \zeta_i^2 + S c_r'^2)\right\} d\zeta_1 \dots d\zeta_m d\sigma. \quad (29)$$

Integrating with respect to all the  $\zeta_i$  except  $\zeta_m$  we have

$$P(d\zeta_m d\sigma | \theta H) \propto \sigma^{-n+m-2} \exp\left\{-\frac{1}{2\sigma^2}(b_m \zeta_m^2 + S c_r'^2)\right\} d\zeta_m d\sigma, \quad (30)$$

and then integrating with regard to  $\sigma$ ,

$$P(d\zeta_m | c_1 \dots c_n H) \propto (S c_r'^2 + b_m \zeta_m^2)^{-1/2(n-m+1)} d\zeta_m \quad (31)$$

$$= \left(\frac{b_m}{\pi S c_r'^2}\right)^{1/2} \frac{\{\frac{1}{2}(n-m-1)\}!}{\{\frac{1}{2}(n-m-2)\}!} \left(1 + \frac{b_m \zeta_m^2}{S c_r'^2}\right)^{-1/2(n-m+1)} d\zeta_m, \quad (32)$$

so that the posterior probability of  $\zeta_m$  is distributed as for  $t$  with  $n-m$  degrees of freedom. It is easily seen that the same applies to any linear function of the  $\zeta_i$ . If  $n-m$  is large the distribution becomes approximately normal with standard error  $\sigma(\zeta_m)$  given by

$$\sigma^2(\zeta_m) = S c_r'^2 / (n-m) b_m = B_{mm} S c_r'^2 / (n-m) D.$$

This is the same as the form taken by (21) if we replace  $\sigma^2$  by

$$Sc_r'^2/(n-m).$$

The practical method of solution is as follows. We start with the  $n$  equations

$$a_{ir}x_i = c_r \quad (33)$$

which are called the *equations of condition*. In general no set of values of  $x_i$  will satisfy them all exactly. But if we multiply each equation by  $a_{jr}$  and sum for all values of  $r$ , we obtain the equation

$$b_{ij}x_i = d_j \quad (34)$$

by the definitions of  $b_{ij}$  and  $d_i$ . This is done for all values of  $j$  from 1 to  $m$ , and yields  $m$  equations for  $x_i$ . These are the *normal equations*. Their solution as simultaneous equations is

$$x_i = y_i. \quad (35)$$

The most convenient process of solution is identical with that of finding the  $\zeta_i$ . For if we divide the first equation by  $b_{11}$  the function on the left is

$$x_1 + \frac{b_{12}}{b_{11}}x_2 + \dots + \frac{b_{1m}}{b_{11}}x_m = \left(y_1 + \frac{b_{12}}{b_{11}}y_2 + \dots + \frac{b_{1m}}{b_{11}}y_m\right) + \zeta_1. \quad (36)$$

Multiplying this in turn by  $b_{12}, b_{13}, \dots$  and subtracting from all the others, we eliminate  $x_1$  from all. Thus we are left with  $m-1$  equations, which still have the property that the coefficient of  $x_i$  in the equation for  $x_j$  is equal to that of  $x_j$  in the equation for  $x_i$ ; for both are equal to

$$b_{ij} - b_{1i}b_{1j}/b_{11}.$$

We can therefore proceed to eliminate all in turn, finishing with  $x_m$ , the coefficient of which will be  $b_m$ , and  $b_m$  is therefore yielded automatically. Any other coefficient  $b_i$  is the coefficient of  $x_i$  in the first equation remaining when  $x_1$  to  $x_{i-1}$  have been eliminated. Thus the process of solution yields all the  $b_i$ . If  $\sigma$  is initially known, all that remains is to express any unknown, say  $x_1$ , in the form  $d_1/b_1 \pm \sigma/\sqrt{b_1} +$  a linear function of  $y_2$  to  $y_m$  and of  $\zeta_2$  to  $\zeta_m$ ; in this we use the second equation to replace  $y_2$  by a constant  $\pm \sigma/\sqrt{b_2}$  with functions of  $y_3$  to  $y_m$  and of  $\zeta_3$  to  $\zeta_m$ , and so on. Thus finally we obtain the value of  $y_1$ , which is the most probable value of  $x_1$ , and a set of independent uncertainties of  $x_1$ , which are easily combined.

If  $\sigma$  is initially unknown we proceed to estimate the  $y_i$  as before; then substituting in the equations of condition we obtain the set of differences  $c_r - a_{ir}y_i$ , which are called the *residuals*, and are identical with  $c'_r$ . Then we can define the standard deviation of one observation by

$$(n-m)s^2 = Sc_r'^2, \quad (37)$$

and that of  $z_m$  by

$$s_{z_m} = s/\sqrt{b_m}. \quad (38)$$

Put  $t = z_m/s_{z_m}$ ; and we have

$$P(dz_m | c_1 \dots c_n H) \propto \left\{ 1 + \frac{b_m s_{z_m}^2 t^2}{(n-m)s^2} \right\}^{-1/2(n-m+1)} dt = \left( 1 + \frac{t^2}{n-m} \right)^{-1/2(n-m+1)} dt,$$

which is of exactly the same form as 3.44 (13). If  $n-m = \nu$ ,  $\nu$  is again the number of degrees of freedom, and the  $t$  table can be used as in the simpler cases.

This method (essentially Gauss's method of substitution) has great advantages over some of those usually given, which involve the working out of  $m+1$  determinants of the  $m$ th order to obtain the  $y_i$ , and the evaluation of the first minors of all terms in the leading diagonal of  $D$  to find the standard errors of the  $y_i$ . Personally I find that to get the right value for a determinant above the third order is usually beyond my powers, but the above process usually gives me the right answer. The symmetry of the equations at each stage of the solution gives a useful check on the arithmetic, and the correctness of the final solution can be checked by substitution.

A method due to Laplace is often said to be independent of the normal law; but it assumes that the 'best' estimate is a linear function of the observations, and if there was only one unknown this would imply by symmetry the postulate of the arithmetic mean, which in turn implies the normal law. Further, it assumes that the error is estimated by the expectation of its square, which is justified by the normal law but has to be taken as a separate (and wrong) postulate otherwise; and an unnecessary appeal to Bernoulli's theorem has to be made.†

**3.51.** To illustrate the method of solution, consider the following set of normal equations (1), (2), (3); the standard deviation of one observation is  $s$ .

$$\begin{array}{ll} 12x - 5y + 4z = 2 & (1) \\ -5x + 8y + 2z = 1 & (2) \\ 4x + 2y + 6z = 5 & (3) \end{array} \quad \left| \begin{array}{l} x - 0.42y + 0.33z = +0.17 \pm 0.28s \\ 5x - 2.1y + 1.7z = +0.8 \\ 4x - 1.7y + 1.3z = +0.7 \end{array} \right. \quad \begin{array}{l} (4) \\ (5) \\ (6) \end{array}$$

$$\begin{array}{ll} 5.9y + 3.7z = +1.8 & (7) \\ 3.7y + 4.7z = +4.3 & (8) \end{array} \quad \left| \begin{array}{l} y + 0.63z = +0.31 \pm 0.41s \\ 3.7y + 2.3z = +1.1 \end{array} \right. \quad \begin{array}{l} (9) \\ (10) \end{array}$$

$$2.4z = +3.2 \quad (11) \quad \left| \begin{array}{l} z = +1.33 \pm 0.64s \end{array} \right. \quad (12)$$

$$y = +0.31 - 0.63 \times 1.33 = -0.53 \quad (13)$$

$$x = +0.17 - 0.42 \times 0.53 - 0.33 \times 1.33 = -0.49 \quad (14)$$

(4) is got by *dividing* (1) by 12; (5), (6) by multiplying (4) by 5 and 4. Then (2) and (5) give (7), and so on. These results should be checked

† Cf. *Phil. Mag.* 22, 1936, 337-59.

by substitution in the original equations. The standard error  $0.28s$  in the first line is  $s/\sqrt{12}$ , and similarly for the others. For  $s_y$  we have

$$s_y = \pm 0.41s \pm 0.63 \times 0.64s = (\pm 0.41 \pm 0.41)s = \pm 0.58s, \quad (15)$$

and for  $s_x$

$$x = (x - 0.42y + 0.33z) + 0.42(y + 0.63z) - 0.60z; \quad (16)$$

$$s_x = (\pm 0.28 \pm 0.42 \times 0.41 \pm 0.60 \times 0.64)s; \quad (17)$$

$$s_x^2 = 0.29s^2, \quad s_x = 0.54s. \quad (18)$$

Hence

$$x = -0.49 \pm 0.54s; \quad y = -0.53 \pm 0.58s; \quad z = +1.33 \pm 0.64s. \quad (19)$$

### 3.52. Equations of condition of unequal weights; Grouping.

In the argument of 3.5 we have assumed that every measure has the same standard error. If the standard errors are unequal, 3.5 (3) will be replaced by

$$P(\theta | x_i, \sigma_r, H) = \frac{1}{(2\pi)^{1/2n} \prod \sigma_r} \exp \left\{ -S \frac{1}{2\sigma_r^2} (c_r - f_r)^2 \right\} \prod dc_r \quad (1)$$

and the exponent is still a quadratic form. It differs from  $W$  in so far as each term of the sum has to be divided by  $\sigma_r^2$  before addition. Consequently the quantities  $\sigma_r^{-2}$ , or their products by a convenient constant, are called the *weights* of the equations of condition. It will be noticed that (1) is the same as if we replaced the equations

$$f_r = c_r \pm \sigma_r \quad (2)$$

$$\text{by} \quad \frac{\sigma f_r}{\sigma_r} = \frac{\sigma c_r}{\sigma_r} \pm \sigma \quad (3)$$

and took each observation as one of  $\sigma f_r / \sigma_r$  with the same uncertainty  $\sigma$ . If the  $\sigma_r$  are known and  $\sigma$  is chosen conveniently the formation and solution of the normal equations will proceed exactly as before. Evidently the arbitrary  $\sigma$  will cancel in the course of the work. This procedure is convenient as an aid to seeing that the method needs only a slight alteration at the outset, and is sometimes recommended as a practical method; that is, it is proposed that the whole of the equations of condition should be multiplied by their respective  $\sigma/\sigma_r$  before forming the normal equations. This has the disadvantage that the weights are often integers and the multiplication brings in square roots and consequent additional rounding-off errors. It is better to proceed as follows. If

$$W = \frac{1}{2} S \left\{ \frac{\sigma}{\sigma_r} (a_{ir} x_i - c_r) \frac{\sigma}{\sigma_r} (a_{jr} x_j - c_r) \right\} \quad (4)$$

$$W \text{ is also equal to } \frac{1}{2} S(a_{ir} x_i - c_r) \left\{ \frac{\sigma^2}{\sigma_r^2} (a_{jr} x_j - c_r) \right\}, \quad (5)$$

$$\text{and } \frac{\partial W}{\partial x_i} = S a_{ir} \left\{ \frac{\sigma^2}{\sigma_r^2} (a_{jr} x_j - c_r) \right\}. \quad (6)$$

Consequently, if we first multiply every equation of condition by its weight  $\sigma^2/\sigma_r^2$ , and then form the normal equations by multiplying by  $a_{ir}$  and adding, we get the same equations with less trouble and more accuracy.

If the  $\sigma_r$  are unknown and some of them mutually irrelevant there will be a complication similar to that of 3.41 (14). But it often happens in a programme of observation that some observations are recorded as made in specially favourable conditions, some moderate, and some poor. It is usual to deal with this by attaching impressions of the relative accuracy in the form of weights, somewhat arbitrarily, though a determination of the accuracy of observations in the various grades would be possible if the residuals were classified. Our problem, if the relative accuracies are accepted, is to obtain an estimate of accuracy when the  $\sigma_r$  are not taken as known, but their ratios are taken as known. We take  $\sigma$  as the standard error corresponding to unit weight and proceed as just described. If  $w_r = \sigma^2/\sigma_r^2$  is the weight of the  $r$ th observation the term  $S c_r'^2$  in 3.5 (29) will be replaced by  $S w_r c_r'^2$ . The only change in the method of estimating  $\sigma$  is therefore that in forming  $s^2$  as in 3.5 (37) we must multiply each  $c_r'^2$  by the weight of the observation.

The observations often fall into groups such that within each group all the  $a_{ir}$  are nearly the same. The extreme case of this condition is the problem of 3.44, where for the  $i$ th station  $a_{ir} = 1$  if the observation is at that station and 0 if it is at any other. In the determination of an earthquake epicentre from the times of arrival of a phase at different stations, the stations fall into geographical regions such that within any region the time of arrival would be altered by nearly the same amount by any change in the adopted time of occurrence and the position of the epicentre. It then simplifies the work considerably to form an equation of condition for the mean position of the stations in the region and to use the mean  $c_r$  for it. The standard error of the latter will be  $\sigma/\sqrt{n_r}$ , where  $n_r$  is the number of stations in the region, and therefore it supplies an equation of condition of weight  $n_r$ . The normal equations will be nearly the same as if all the stations were used to form separate equations of condition. All the residuals are still available to provide an estimate of  $\sigma$ , which will be on  $n-m$  degrees of freedom

just as in the treatment without grouping. If we chose to use the method described in the last paragraph we should get the same least squares solution, but only the mean residuals in the groups would be available to provide an estimate of uncertainty, which would therefore be on many fewer degrees of freedom.

3.53. The following data, given by E. C. Bullard and H. L. P. Jolly,† provide a more complicated instance of the method. The unknowns are the values of gravity at various places. In general gravity is not measured absolutely, but the difference between the periods of the same pendulum when swung in different places is found, thus giving an estimate of the difference of gravity. This is referred to a standard value for Potsdam, where an absolute determination exists. In the following set of equations of condition, therefore, absolute values refer to stations compared directly with Potsdam; the rest are differences. Bullard and Jolly took De Bilt as given, but it appears that the comparison of De Bilt with Potsdam has an appreciable uncertainty compared with those of some of the English stations, and it seems best to treat it as an additional unknown. The unknowns are then:

- $g_0$ , De Bilt.
- $g_1$ , Greenwich, Record Room.
- $g_2$ , Greenwich, National Gravity Station.
- $g_3$ , Kew.
- $g_4$ , Cambridge, Pendulum House.
- $g_5$ , Southampton.

The equations of condition are:

<i>Observer</i>	<i>Date</i>		
Putnam . . . . .	1900	$g_1 = 981.188$	(1)
Putnam . . . . .	1900	$g_2 = 981.200$	(2)
Lenox-Conyngham . . . . .	1903	$g_2 - g_1 = +0.014$	(3)
Meinesz . . . . .	1925	$g_4 - g_0 = -0.003$	(4)
Lenox-Conyngham and Manley . . . . .	1925	$g_4 - g_2 = +0.0647$	(5)
Jolly and McCaw . . . . .	1927	$g_2 - g_1 = +0.0003$	(6)
Miller . . . . .	1928	$g_2 = 981.1888$	(7)
Jolly and Willis . . . . .	1930	$g_4 - g_2 = +0.0742$	(8)
Willis and Bullard . . . . .	1931	$g_2 - g_1 = +0.0653$	(9)
Jolly and Bullard . . . . .	1933	$g_4 - g_2 = +0.1431$	(10)
Bullard . . . . .	1935	$g_4 - g_2 = +0.1390$	(11)
Meinesz . . . . .	1921	$g_0 = 981.267$	(12)
Meinesz . . . . .	1925	$g_0 = 981.269$	(13)

The unit is 1 gal = 1 cm./sec.<sup>2</sup>

A main source of error is known to be change of the mechanical properties of the pendulums during transport. Hence all the equations will

† *M.N.R.A.S.*, Geophys. Suppl. 3, 1936, 470.

be taken of equal weight except (6). For this the stations are only 300 metres apart and at nearly the same height, and the difference can be calculated more accurately than it can be measured. I take

$$g_2 - g_1 = +0.0001.$$

An approximate set of solutions is easily found; we write

$$g_0 = 981.268 + x_0, \quad (14)$$

$$g_1 = 981.188 + x_1, \quad (15)$$

$$g_2 = 981.1881 + x_1, \quad (16)$$

$$g_3 = 981.200 + x_3, \quad (17)$$

$$g_4 = 981.265 + x_4, \quad (18)$$

$$g_5 = 981.123 + x_5. \quad (19)$$

Then the equations of condition, omitting (6), become

$$x_1 = 0.0000, \quad (1')$$

$$x_3 = 0.0000, \quad (2')$$

$$x_3 - x_1 = +0.0020, \quad (3')$$

$$x_4 - x_0 = 0.0000, \quad (4')$$

$$x_4 - x_3 = -0.0003, \quad (5')$$

$$x_1 = +0.0007, \quad (7')$$

$$x_4 - x_1 = -0.0027, \quad (8')$$

$$x_1 - x_5 = +0.0002, \quad (9')$$

$$x_4 - x_5 = +0.0011, \quad (10')$$

$$x_4 - x_5 = -0.0030, \quad (11')$$

$$x_0 = -0.0010, \quad (12')$$

$$x_0 = +0.0010. \quad (13')$$

$x_0$  occurs in equations (4'), (12'), (13'), with coefficients  $-1, +1, +1$ . We therefore add (12') and (13') and subtract (4') to give the normal equation for  $x_0$ , namely

$$3x_0 - x_4 = 0.0000.$$

$x_1$  occurs in (1'), (7'), (9') with coefficient  $+1$ , in (3'), (8') with coefficient  $-1$ . We therefore multiply (3'), (8') by  $-1$  and add to the sum of (1'), (7'), (9'). Similarly we proceed for the others.

#### Normal equations

$$\begin{array}{rcll} 3x_0 & -x_4 & = & 0.0000 \quad (20) \\ 5x_1 - x_3 - x_4 - x_5 & = & +0.0016 & (21) \\ -x_1 + 3x_3 - x_4 & = & +0.0023 & (22) \\ -x_0 - x_1 - x_3 + 5x_4 - 2x_5 & = & -0.0049 & (23) \\ -x_1 & -2x_4 + 3x_5 & = & +0.0017 \quad (24) \end{array} \quad \left| \begin{array}{l} x_0 \\ x_1 \\ x_3 \\ x_4 \\ x_5 \end{array} \right. \quad \begin{array}{l} -0.3333x_4 = 0.0000 \quad (25) \end{array}$$

First divide (20) by 3; the result is (25). To eliminate  $x_0$  we have only to add (25) to (23). Then

$$\begin{array}{lcl} 5x_1 - x_3 & -x_4 - x_5 = +0.0016 & (21) \\ -x_1 + 3x_3 & -x_4 & = +0.0023 \quad (22) \\ -x_1 - x_3 + 4.6667x_4 - 2x_5 & = -0.0049 & (26) \\ -x_1 & -2x_4 + 3x_5 = +0.0017 & (24) \end{array} \quad \left| \quad \begin{array}{l} x_1 - 0.2x_3 - 0.2x_4 - 0.2x_5 = +0.00032 \quad (27) \end{array} \right.$$

Now eliminate  $x_1$ ;

$$\begin{array}{lcl} 2.8x_3 & -1.2x_4 - 0.2x_5 = +0.00262 & (28) \\ -1.2x_3 + 4.4667x_4 - 2.2x_5 & = -0.00458 & (29) \\ -0.2x_3 & -2.2x_4 - 2.8x_5 = +0.00202 & (30) \\ 3.9524x_4 - 2.2857x_5 & = -0.00345 & (34) \\ -2.2857x_4 + 2.7857x_5 & = +0.00221 & (35) \\ 1.4639x_5 & = +0.00021 & (38) \end{array} \quad \left| \quad \begin{array}{l} x_2 - 0.4286x_4 - 0.0714x_5 = +0.00094 \quad (31) \\ 1.2x_3 - 0.5143x_4 - 0.0857x_5 = +0.00113 \quad (32) \\ 0.2x_3 - 0.0857x_4 - 0.0143x_5 = +0.00019 \quad (33) \\ x_4 - 0.5783x_5 = -0.00087 \quad (36) \\ 2.2857x_4 - 1.3218x_5 = -0.00200 \quad (37) \\ x_5 = +0.00014 \quad (39) \end{array} \right.$$

Hence the solution is, from (36), (31), (27), (25) in turn,

$$\left. \begin{array}{l} x_0 = -0.00020 \\ x_1 = +0.00031 \\ x_3 = +0.00061 \\ x_4 = -0.00079 \\ x_5 = +0.00014 \end{array} \right\} \quad (40)$$

Substituting in the normal equations we find that the largest discrepancy is 5 in the fifth decimal, so that the solution is checked. A check on the formation of the normal equations is got by noticing that most of the equations of condition are differences; hence the sum of the right sides of (1'), (2'), (7'), (12'), (13') should be that of the right sides of (20) to (24).† Now substituting in the equations of condition we get the calculated values. Residuals are multiplied by 1000 for convenience.

	Calc.	O - C	c' <sup>2</sup>
(1')	+0.31	-0.31	0.10
(2')	+0.61	-0.61	0.37
(3')	+0.30	+1.70	2.89
(4')	-0.59	+0.59	0.35
(5')	-1.40	+1.10	1.21
(7')	+0.31	+0.39	0.15
(8')	-1.10	-1.60	2.56
(9')	+0.17	+0.03	0.01
(10')	-0.93	+2.03	4.12
(11')	-0.93	-2.07	4.28
(12')	-0.20	-0.80	0.64
(13')	-0.20	+1.20	1.44
			18.12

† For general methods of checking when the number of normal equations is large, see H. and B. S. Jeffreys, *Methods of Mathematical Physics*, p. 283. Method (2) mentioned on p. 284 will also check the formation of the normal equations themselves from the equations of condition.



We have 12 equations and 5 unknowns have been found; hence

$$s^2 = 18.12/(12-5) = 2.59; \quad s = 1.60 \text{ milligal.} \quad (41)$$

The uncertainties of the separate determinations have still to be found. Denote departures from the least squares solutions by accents and take 1 milligal as the unit. Denote, apart from accents, the quantities on the right of (25), (27), (31), (36), (39) by  $X_i$ ; these have independent uncertainties. Then

$$X'_5 = x'_5 = \pm 1.60/(1.46)^{1/2} = \pm 1.32, \quad (42)$$

$$X'_4 = \pm 1.60/(3.95)^{1/2} = \pm 0.80, \quad (43)$$

$$X'_3 = \pm 1.60/(2.80)^{1/2} = \pm 0.96, \quad (44)$$

$$X'_1 = \pm 1.60/(5.00)^{1/2} = \pm 0.71, \quad (45)$$

$$X'_0 = \pm 1.60/(3.00)^{1/2} = \pm 0.92. \quad (46)$$

But

$$x'_4 = X'_4 + 0.578x'_5 = \pm 0.80 \pm 0.76 = \pm 1.10, \quad (47)$$

$$x'_3 = X'_3 + 0.4286X'_4 + 0.177x'_5, \quad (48)$$

$$= \pm 0.96 \pm 0.34 \pm 0.23 = \pm 1.04. \quad (49)$$

and so on. The final solution is

$$g_0 = 981.26780 \pm 0.00099, \quad (50)$$

$$g_1 = 981.18831 \pm 0.00092, \quad (51)$$

$$g_2 = 981.18841 \pm 0.00092, \quad (52)$$

$$g_3 = 981.20061 \pm 0.00104, \quad (53)$$

$$g_4 = 981.26421 \pm 0.00110, \quad (54)$$

$$g_5 = 981.12314 \pm 0.00132. \quad (55)$$

From the  $t$  table for 7 degrees of freedom we find that the probability of an error numerically greater than 2 milligals ranges from about 0.07 for  $g_1$  and  $g_2$  to 0.18 for  $g_5$ .

The standard errors are not much less than for one determination. This is ultimately because, of the 12 equations, only 5 represent direct comparisons with Potsdam. Even if the differences were exactly determined the standard errors could not be less than  $1.60/\sqrt{5} = 0.72$  milligal. The fact that most of the equations give differences makes the normal equations far from orthogonal, as is shown by the fact that the coefficient of  $x_5$  drops from 3.0 in (24) to 1.46 in (38).

Seidel's method (see p. 173) was tried on these equations, but convergence was too slow. This method is really adapted only to problems where the equations are nearly orthogonal, otherwise the estimation of uncertainty becomes more laborious than the solution of the normal equations.

With a slight modification, however, the method succeeds. The difficulty arises principally from  $x_5$ ; the equations give direct determinations of  $x_0$ ,  $x_1$ , and  $x_3$ , while  $x_4$  is connected directly to all these three.

But  $x_5$  has only a single connexion with  $x_1$  and two with  $x_4$ . Hence  $x_5$  really has little to say concerning the values of the other four, which would be well determined without it. If we drop the equations containing  $x_5$  we have the normal equations

$$3x_0 \quad \quad -x_4 = 0.0000, \quad (56)$$

$$4x_1 - x_3 - x_4 = +0.0014, \quad (57)$$

$$-x_1 + 3x_3 - x_4 = +0.0023, \quad (58)$$

$$-x_0 - x_1 - x_3 + 3x_4 = -0.0030. \quad (59)$$

These are nearly orthogonal. The largest term on the right is in (59); we therefore take a first approximation  $x_4 = -0.0010$ . Then from (56),  $x_0 = -0.0003$ ; from (58),  $x_3 = +0.0004$ ; and from (57),  $x_1 = +0.0002$ . Substituting these approximations in the left sides we have in turn

$$3x_0 \quad \quad -x_4 = +0.0001,$$

$$4x_1 - x_3 - x_4 = +0.0014,$$

$$-x_1 + 3x_3 - x_4 = +0.0020,$$

$$-x_0 - x_1 - x_3 + 3x_4 = -0.0033.$$

Comparing with the original equations we see that (58) and (59) are both  $+0.0003$  higher, and that we can add  $0.0001$  to  $x_3$  and  $x_4$ . Then

$$x_0 = -0.0003; \quad x_1 = +0.0002; \quad x_3 = +0.0005; \quad x_4 = -0.0009.$$

This is very near the solution (40).

If the equations were strictly orthogonal the standard errors would be  $\sigma/\sqrt{3}$ ,  $\sigma/2$ ,  $\sigma/\sqrt{3}$ ,  $\sigma/\sqrt{3}$ , and independent. To a second approximation

$$\sigma^2(x_0) = \frac{1}{3}\sigma^2 + \frac{1}{6}\sigma^2(x_4),$$

$$\sigma^2(x_1) = \frac{1}{4}\sigma^2 + \frac{1}{18}\sigma^2(x_3) + \frac{1}{18}\sigma^2(x_4),$$

$$\sigma^2(x_3) = \frac{1}{3}\sigma^2 + \frac{1}{6}\sigma^2(x_1) + \frac{1}{6}\sigma^2(x_4),$$

$$\sigma^2(x_4) = \frac{1}{3}\sigma^2 + \frac{1}{6}\sigma^2(x_0) + \frac{1}{6}\sigma^2(x_1) + \frac{1}{6}\sigma^2(x_3).$$

By iteration we find, nearly,

$$\sigma^2(x_0) = 0.40\sigma^2; \quad \sigma^2(x_1) = 0.31\sigma^2; \quad \sigma^2(x_3) = 0.42\sigma^2; \quad \sigma^2(x_4) = 0.46\sigma^2.$$

Again,

$$\begin{aligned} x_5 &= \frac{2}{3}(x_4 + 0.0009) + \frac{1}{3}(x_1 - 0.0002) \pm \sigma/\sqrt{3} \\ &= 0.0000 \pm 0.56^{1/2}\sigma. \end{aligned}$$

$\sigma$  is estimated as before, and the solution is

$$x_0 = -0.0003 \pm 0.0010,$$

$$x_1 = +0.0002 \pm 0.0009,$$

$$x_3 = +0.0005 \pm 0.0010,$$

$$x_4 = -0.0009 \pm 0.0011,$$

$$x_5 = 0.0000 \pm 0.0012.$$

The accuracy would be enough for all practical purposes.

**3.54.** The following problem, and various extensions of it, have often occurred in astronomy. There are cases where a group of stars can be assumed all to have the same parallax; the estimates from any star separately are comparable with their standard errors, but the mean of all is substantially more than its standard error. The physical restriction here is that a parallax cannot be negative. It is substantially less than the standard error of one observation, and we may adopt a uniform prior probability over positive values. If, then,  $\alpha$  is the general parallax and  $a_r, s_r$  the separate estimates with their standard errors, the number of observations in each case being large, we have

$$P(da_1 \dots da_n | \alpha H) \propto \exp\left\{-\sum \frac{(\alpha - a_r)^2}{2s_r^2}\right\} da_1 \dots da_n$$

$$\text{and} \quad P(d\alpha | H) \propto d\alpha \quad (\alpha > 0); \quad = 0 \quad (\alpha < 0).$$

Then

$$P(d\alpha | a_1 \dots a_n H) \propto \exp\left\{-\sum \frac{(\alpha - a_r)^2}{2s_r^2}\right\} d\alpha \quad (\alpha > 0); \quad = 0 \quad (\alpha < 0).$$

The posterior probability of  $\alpha$  is therefore a normal one about the weighted mean of the  $a_r$ , but it is truncated at  $\alpha = 0$ .

The treatment of such problems has given rise to some discussion. In the conditions of the problem some of the estimates  $a_r$  are usually negative. These have sometimes been rejected as impossible, and a mean is taken of the positive ones. Then the rejection of a large fraction of the negative random errors biases the mean by an amount comparable with the standard error of one determination. We are entitled to allow for the impossibility of a negative true parallax, but this can only be done at the end when we take the prior probability into account. If only one star was in question we should still be entitled to take it into account. We must not, however, do it by rejecting factors from the likelihood. The point is somewhat similar to one that arises in one case of the combination of correlation coefficients (p. 156), where there is a constant term in  $\zeta - z$  arising partly from the prior probability and partly from the likelihood. But when several estimates are combined the part from the prior probability only enters once, while that from the likelihood enters every time. Similar considerations have occurred in the estimation of the focal depths of shallow earthquakes. Here the depth  $h$  enters through  $h^2$ ; and the least squares solution is liable to give negative  $h^2$ . There are two valid treatments possible. One is to take  $h$  as zero in all cases in the estimation of other parameters, especially the velocities, thus regarding the whole of the estimated values as not

significant. The other is to eliminate  $h$  from all the solutions and combine the equations for the velocities. What is not valid is to reject the cases of negative estimated  $h^2$  and determine the velocities from the rest; this gives a bias in the estimated velocities.

**3.6. The rectangular distribution.** This distribution is of theoretical interest on account of the fact that the mean of all the observed values gives a less accurate estimate of the centre of the distribution than the mean of the two extreme observations does by itself. Let the centre of the distribution be  $\alpha$  and the range  $2\sigma$ , to be determined. The chance of an observation in a range  $dx$  is

$$P(dx | \alpha, \sigma, H) = \begin{cases} dx/2\sigma & (\alpha - \sigma < x < \alpha + \sigma), \\ 0 & (x < \alpha - \sigma, x > \alpha + \sigma). \end{cases} \quad (1)$$

$$(2)$$

The chance of  $n$  observations in given ranges is

$$P(dx_1 \dots dx_n | \alpha, \sigma, H) = \prod (dx)/(2\sigma)^n, \quad (3)$$

provided that all the  $x_r$  satisfy the conditions

$$\alpha - \sigma < x_r < \alpha + \sigma, \quad (4)$$

and therefore provided that the extreme observations satisfy them. Call these  $x_1$  and  $x_2$ . We take  $\alpha$  and  $\sigma$  as initially unknown, and therefore

$$P(d\alpha d\sigma | H) \propto d\alpha d\sigma / \sigma \quad (5)$$

and

$$P(d\alpha d\sigma | x_1 \dots x_n H) \propto \sigma^{-n-1} d\alpha d\sigma, \quad (6)$$

provided now

$$\alpha - \sigma < x_1; \quad \alpha + \sigma > x_2. \quad (7)$$

These conditions fix the possible joint range of  $\alpha$  and  $\sigma$ , given the observations, and apart from the restrictions on the range the observations do not appear in (6). Hence, with the rectangular law, the two extreme observations are sufficient statistics for  $\alpha$  and  $\sigma$ .

Then 
$$P(d\alpha | x_1 \dots x_n H) \propto d\alpha \int \sigma^{-n-1} d\sigma \quad (8)$$

through the permitted range. But, given  $\alpha$ ,  $\sigma$  must be greater than the larger of  $\alpha - x_1$  and  $x_2 - \alpha$ ; the lower limit for  $\sigma$  is therefore  $\alpha - x_1$  if  $\alpha > \frac{1}{2}(x_1 + x_2)$ , and  $x_2 - \alpha$  if  $\alpha < \frac{1}{2}(x_1 + x_2)$ . Hence

$$P(d\alpha | x_1 \dots x_n H) \propto \begin{cases} (\alpha - x_1)^{-n} d\alpha & \{\alpha > \frac{1}{2}(x_1 + x_2)\}, \\ (x_2 - \alpha)^{-n} d\alpha & \{\alpha < \frac{1}{2}(x_1 + x_2)\}, \end{cases} \quad (9)$$

$$(10)$$

with the same constant factor in both cases. The posterior probability for  $\alpha$ , therefore, has a sharp peak at the mean of the extreme values.

The constant factor is easily found to be  $2^{-n}(n-1)(x_2-x_1)^{n-1}$ . If  $n = 2$ , we have

$$\begin{aligned} P(x_1 < \alpha < x_2 | x_1, x_2, H) \\ &= 2^{-n}(n-1)(x_2-x_1)^{n-1} \left\{ \int_{x_1}^{1/2(x_1+x_2)} (x_2-\alpha)^{-n} d\alpha + \int_{1/2(x_1+x_2)}^{x_2} (\alpha-x_1)^{-n} d\alpha \right\} \\ &= \frac{1}{2}. \end{aligned} \quad (11)$$

Thus, if we have only two observations, and  $\alpha$  and  $\sigma$  are originally unknown, the posterior probability that  $\alpha$  lies between the observed values is  $\frac{1}{2}$ . This is a general rule for any continuous law of error; we have already had a case of it for the normal law.

The possible values of  $\alpha$ , given  $\sigma$ , range from  $x_2-\sigma$  to  $x_1+\sigma$ , provided the latter is the greater. Then

$$P(d\sigma | x_1 \dots x_n H) \propto \frac{x_1-x_2+2\sigma}{\sigma^{n+1}} d\sigma \quad (12)$$

for  $\sigma > \frac{1}{2}(x_2-x_1)$ . The constant factor is found to be

$$2^{-n}n(n-1)(x_2-x_1)^{n-1}.$$

If  $n = 1$ , the range for  $\alpha$  is from  $x_2-\sigma$  to  $x_1+\sigma$ , and (6) leads to

$$P(d\sigma | x_1, H) \propto d\sigma/\sigma, \quad (13)$$

which expresses the same fact as for the normal law, that one observation can tell us nothing about its own accuracy. It may be noticed that the probability density for  $\sigma$  vanishes at  $\sigma = \frac{1}{2}(x_2-x_1)$  and has a maximum at  $\sigma = \frac{1}{2}(1+1/n)(x_2-x_1)$ . This is because the extreme value would require both  $x_1$  and  $x_2$  to have fallen at the extremes of the law, which would be surprising, but it would not be surprising that both should fall a little within them.

On account of the form of the limiting conditions the posterior probabilities of  $\alpha$  and  $\sigma$  are far from independent; any inference that involves both should proceed from (6) directly. If we want the termini  $\alpha_1 = \alpha - \sigma$  and  $\alpha_2 = \alpha + \sigma$ , (6) transforms to

$$P(d\alpha_1 d\alpha_2 | x_1 \dots x_n H) \propto d\alpha_1 d\alpha_2 / (\alpha_2 - \alpha_1)^{n+1} \quad (\alpha_1 < x_1, \alpha_2 > x_2); \quad (14)$$

whence for  $\alpha_2 > x_2$ ,

$$P(d\alpha_2 | x_1 \dots x_n H) = (n-1)(x_2-x_1)^{n-1}(\alpha_2-x_1)^{-n} d\alpha_2. \quad (15)$$

If we fix limits such that the probability that  $\alpha$ ,  $\alpha_1$ , or  $\alpha_2$  lies between them has any definite value, the distance between these limits will decrease like  $1/n$  as the number of observations increases, whereas with the normal law of error the corresponding distance decreases like  $1/\sqrt{n}$ . This kind of result usually arises for laws of error with a finite range

where the gradient of the law is non-zero at an extreme, and especially for any U-shaped or J-shaped law. The rectangular law is merely the transition from a bell-shape to a U-shape.

The use of the mean and second moment as location and scale parameters in such cases sacrifices much information. For with the rectangular law the second moment of the law is  $\frac{1}{3}\sigma^2$ , and the standard error of the mean of  $n$  observations, given  $\sigma$ , will be  $\sigma/\sqrt{(3n)}$ , thus diminishing like  $1/\sqrt{n}$ , whereas any range for a definite probability that  $\alpha$  lies within it will diminish like  $1/n$  if we use the most accurate methods of fitting.

**3.61. Re-scaling of a law of chance.** As many laws do not lead to sufficient statistics, as the normal and rectangular laws do, it has sometimes been suggested that it would be beneficial to choose a new variable whose law will be normal or rectangular. Thus if the law is

$$P(dx | \alpha, \sigma, H) = f\left(\frac{x-\alpha}{\sigma}\right) \frac{dx}{\sigma},$$

we can define

$$\frac{y-\alpha}{\sigma} = \int_{-\infty}^x f\left(\frac{x-\alpha}{\sigma}\right) \frac{dx}{\sigma},$$

and then  $P(dy | \alpha, \sigma, H) = dy/\sigma \quad (\alpha < y < \alpha + \sigma).$

Similarly we could define a  $z$  such that

$$\frac{1}{\sqrt{(2\pi)\sigma}} \int_{-\infty}^z \exp\left\{-\frac{(z-\beta)^2}{2\sigma^2}\right\} dz = \int_{-\infty}^x f\left(\frac{x-\alpha}{\sigma}\right) \frac{dx}{\sigma},$$

and the chance of  $z$  is normally distributed about  $\beta$  with standard error  $\sigma$ .

It has been suggested that such transformations can be used to simplify methods of estimation, but they are useless. In the first place, for given  $x$  we do not know the corresponding value of  $y$  or  $z$  until we know  $\alpha$  and  $\sigma$ ; and the whole reason for an estimation problem is that we do not. In the second, if  $x$  can be transformed so that

$$\int_{-\infty}^x f(x) dx = \int_{-\infty}^y g(y) dy,$$

then  $P(dy | \alpha, \sigma, H) = g(y) dy = f(x) \frac{dx}{dy} dy,$

where  $dx/dy$  will also depend on  $\alpha$  and  $\sigma$ . If values of  $x$  are observed, the correct likelihood factor is  $\prod f(x_r)$ . But if instead we use  $y$  we shall get a factor  $\prod g(y_r)$ . Thus the two likelihoods will differ by a factor  $\prod (dy/dx)_{x=x_r}$ , a function depending on  $\alpha$  and  $\sigma$  for every observation.

It is remarkable that such maltreatment of the likelihood is recommended (but so far as I know not used because it cannot be) by statisticians who object to the prior probability, which only appears once in any given problem.

**3.62. Reading of a scale.** The commonest case where errors do not satisfy a normal law is the measurement of a length by means of a scale, the positions of the ends being read to the nearest multiple of the scale interval. Let the length of the object be  $L$  units. Two cases arise. In the first, we place one end of the object at a graduation, say the  $m$ th, and read the position of the other to the nearest graduation. Then clearly we shall always record the length as  $k$  units, where  $k$  is the integer nearest to  $L$ . Hence, for any  $k$ ,

$$P(k | LH) = 1 \quad (-\frac{1}{2} < L - k < \frac{1}{2}), \quad P(k | LH) = 0 \quad (|L - k| \geq \frac{1}{2}).$$

If  $n$  observations are made, and  $P(dL | H) \propto dL$ ,

$$P(dL | \theta H) = dL \quad (-\frac{1}{2} < L - k < \frac{1}{2}),$$

$$P(dL | \theta H) = 0 \quad (|L - k| > \frac{1}{2}).$$

In this simple case increasing the number of measurements does nothing to increase the accuracy of the determination. The posterior probability distribution is rectangular.

In the second case, we may put one end at an arbitrary position on the scale, say at  $m + y$  units from one end, where  $-\frac{1}{2} < y < \frac{1}{2}$ ; if the length is  $L = k + x$  units, where  $0 < x < 1$ , the nearest graduation to the other end will be the  $(m + k)$ th if  $|x + y| < \frac{1}{2}$ , that is, if  $-\frac{1}{2} < y < \frac{1}{2} - x$ , and will be the  $(m + k + 1)$ th if  $|x + y| > \frac{1}{2}$ , that is, if  $\frac{1}{2} - x < y < \frac{1}{2}$ . But

$$P(dy | H) = dy \quad (|y| < \frac{1}{2}), \quad P(dy | H) = 0 \quad (|y| > \frac{1}{2})$$

and therefore

$$P(k | LH) = 1 - x; \quad P(k + 1 | LH) = x.$$

If  $r$  observations give the value  $k$ , and  $s$  the value  $k + 1$ , we have

$$P(\theta | LH) = (1 - x)^r x^s;$$

$$P(dL | \theta H) \propto (1 - x)^r x^s dx = \frac{(r + s + 1)!}{r! s!} (1 - x)^r x^s dx.$$

The coefficient of  $dx$  is a maximum if

$$x = \frac{r}{r + s} = x_0$$

so that the most probable value is the mean of the observed values.

For  $r, s$  large the standard error is nearly  $\left\{ \frac{rs}{(r + s)^3} \right\}^{1/2}$ , which is not independent of  $x_0$ .

By a theorem due to Gauss (p. 189), if the probability of an error, given the true value, is a function of the error alone, and if the likelihood is a maximum when the true value is taken to be the mean of the observed values, the law of error must be normal. In this problem the second condition is true but the conclusion is false. The first condition is false because  $P(k | LH)$  is not a function of  $k - L$  alone; if we vary  $L$  but keep  $k - L$  an integer,  $k$  will take non-integral values, which are forbidden by the conditions of the problem. Keynes has shown† that if the law of error is

$$P(dx | \xi H) = f(x, \xi) dx,$$

where  $f$  is twice differentiable with regard to  $\xi$ , a necessary and sufficient condition for the maximum likelihood estimate to be always the arithmetic mean of the observed values  $x_r$  is

$$\log f(x, \xi) = \phi'(\xi)(\xi - x) - \phi(\xi) + \psi(x).$$

The normal law corresponds to

$$\phi(\xi) = -\xi^2/2\sigma^2; \quad \psi(x) = -x^2/2\sigma^2 + \text{constant}.$$

The law for measurement by difference corresponds to

$$\phi(\xi) = (\xi - 1)\log(1 - \xi) - \xi \log \xi; \quad \psi(x) = 0.$$

The Poisson law corresponds to

$$\phi(\xi) = -\xi \log \xi; \quad \psi(x) = -x - \log x!.$$

These reductions to Keynes's form are due to M. S. Bartlett.‡

This problem is of some theoretical interest. In practice the peculiar behaviour of the posterior probability would lead to difficulties in calculation. These are reduced if we can reduce the step of the scale, for instance by means of a microscope, so that the error of reading is no longer the principal source of error.

**3.7. The posterior probabilities that the true value, or the third observation, will lie between the first two observations.** Let us suppose that a law of error is given by  $hf\{h(x - \alpha)\}dx$ , where  $f$  may have any form and  $h$  plays the part of the precision constant, or the reciprocal of the scale parameter. Put

$$\int_{-\infty}^{\infty} f(z) dz = F(z), \quad F(\infty) = 1. \quad (1)$$

† *Treatise on Probability*, p. 197.

‡ *Proc. Roy. Soc. A*, **141**, 1933, 524-5.



If  $\alpha$  and  $h$  are originally unknown, we have

$$P(d\alpha dh | H) \propto d\alpha dh/h, \quad (2)$$

$$P(dx_1 dx_2 | \alpha, h, H) = h^2 f\{h(x_1 - \alpha)\} f\{h(x_2 - \alpha)\} dx_1 dx_2, \quad (3)$$

and  $P(d\alpha dh | x_1, x_2, H) \propto h f\{h(x_1 - \alpha)\} f\{h(x_2 - \alpha)\} d\alpha dh. \quad (4)$

The probability, given  $x_1$  and  $x_2$  ( $x_2 > x_1$ ), that the third observation will lie in any range  $dx_3$ , is

$$P(dx_3 | x_1, x_2, H) = \iint P(dx_3 d\alpha dh | x_1, x_2, H), \quad (5)$$

integrated over all possible values of  $\alpha$  and  $h$ ,

$$\propto dx_3 \iint h^2 f\{h(x_1 - \alpha)\} f\{h(x_2 - \alpha)\} f\{h(x_3 - \alpha)\} d\alpha dh. \quad (6)$$

Let us transform the variables to

$$\theta = h(x_1 - \alpha), \quad \phi = h(x_2 - \alpha). \quad (7)$$

The probability, given  $x_1$  and  $x_2$ , that  $\alpha$  is between them is  $I_1/I_2$ , where  $I_1$  and  $I_2$  are got by integrating (4) from 0 to  $\infty$  with regard to  $h$ , and respectively from  $x_1$  to  $x_2$  and from  $-\infty$  to  $\infty$  with regard to  $\alpha$ . Then

$$(x_2 - x_1)I_1 \propto \int_{-\infty}^0 \int_0^\infty f(\theta)f(\phi) d\theta d\phi = F(0)\{1 - F(0)\}. \quad (8)$$

$$(x_2 - x_1)I_2 \propto \int_{-\infty}^\infty \int_\theta^\infty f(\theta)f(\phi) d\theta d\phi = \int_{-\infty}^\infty f(\theta)\{1 - F(\theta)\} d\theta = 1 - \frac{1}{2} = \frac{1}{2}. \quad (9)$$

Hence  $I_1/I_2 = 2F(0)\{1 - F(0)\}. \quad (10)$

If then  $F(0) = \frac{1}{2}$ , the ratio is  $\frac{1}{2}$ . In all other cases it is less than  $\frac{1}{2}$ . Referring to (1) we see that  $F(0) = \frac{1}{2}$  is the statement that for any given values of  $\alpha$  and  $h$  an observation is as likely to exceed  $\alpha$  as to fall short of it. There will be such a value for any continuous law of given form. Hence, if we define the true value to mean the median of the law, then the probability, given the first two observations, that the true value lies between them is  $\frac{1}{2}$ , whatever their separation. If we chose any other location parameter than the median of the law, and the law was unsymmetrical, the ratio would be less than  $\frac{1}{2}$ . This is a definite reason for choosing the median as the location parameter in any case where the form of the law is unknown. We have already obtained the result in the special cases of the normal and rectangular laws.

The probability that  $x_3$  will lie between  $x_1$  and  $x_2$ , given the latter, is  $I_3/I_4$ , where

$$I_3 = \int_{-\infty}^{\infty} \int_0^{\infty} \int_{x_1}^{x_2} h^2 f\{h(x_1 - \alpha)\} f\{h(x_2 - \alpha)\} f\{h(x_3 - \alpha)\} d\alpha dh dx_3 \quad (11)$$

$$= \int_{-\infty}^{\infty} \int_0^{\infty} h f\{h(x_1 - \alpha)\} f\{h(x_2 - \alpha)\} [F\{h(x_2 - \alpha)\} - F\{h(x_1 - \alpha)\}] d\alpha dh$$

$$= \frac{1}{x_2 - x_1} \int_{-\infty}^{\infty} \int_{\theta}^{\infty} f(\theta) f(\phi) \{F(\phi) - F(\theta)\} d\theta d\phi$$

$$= \frac{1}{x_2 - x_1} \int_{-\infty}^{\infty} [\frac{1}{2} f(\theta) \{1 - F^2(\theta)\} - f(\theta) F(\theta) \{1 - F(\theta)\}] d\theta$$

$$= \frac{1}{x_2 - x_1} (\frac{1}{2} - \frac{1}{6} - \frac{1}{2} + \frac{1}{3}) = \frac{1}{6(x_2 - x_1)}; \quad (12)$$

$$I_4 = \int_{-\infty}^{\infty} \int_0^{\infty} \int_{-\infty}^{\infty} h^2 f\{h(x_1 - \alpha)\} f\{h(x_2 - \alpha)\} f\{h(x_3 - \alpha)\} d\alpha dh dx_3 \quad (13)$$

$$= \int_{-\infty}^{\infty} \int_0^{\infty} h f\{h(x_1 - \alpha)\} f\{h(x_2 - \alpha)\} d\alpha dh$$

$$= \int_{-\infty}^{\infty} \int_{\theta}^{\infty} f(\theta) f(\phi) d\theta d\phi$$

$$= \int_{-\infty}^{\infty} f(\theta) \{1 - F(\theta)\} d\theta$$

$$= \frac{1}{2(x_2 - x_1)}; \quad (14)$$

whence

$$I_3/I_4 = \frac{1}{3}. \quad (15)$$

Thus, if the location and scale parameters are initially unknown, the probability that the third observation will lie between the first two, given the first two, is  $\frac{1}{3}$  whatever the separation of the first two.

The converse theorem to (10) would be that if the posterior probability that the median of the law lies between the first two observations is  $\frac{1}{2}$  whatever their separation, then the prior probability for  $h$  must be  $dh/h$ . If it was  $\lambda(bh)dh/h$ , where  $b$  is a quantity of the dimensions of  $\alpha$  or of  $1/h$ , the ratio  $I_1/I_2$  would involve  $b/(x_2 - x_1)$  and could not be the same for all values of  $x_2 - x_1$ . The only possible modification would therefore be

$$P(d\alpha dh/H) \propto h\gamma^{-1} d\alpha dh. \quad (16)$$

The question is whether  $\gamma$  is necessarily 0 for all admissible forms of  $f(z)$ . If we put

$$h\{\frac{1}{2}(x_2+x_1)-\alpha\} = t, \quad (17)$$

$$\frac{1}{2}h(x_2-x_1) = s \quad (18)$$

$$\text{we find } hf\{h(x_1-\alpha)\}f\{h(x_2-\alpha)\}d\alpha = -f(t-s)f(t+s)dt, \quad (19)$$

and (16) in place of (2) will lead to

$$I_1 - \frac{1}{2}I_2 \propto \int_0^\infty \left( 2 \int_{-s}^s - \int_{-\infty}^\infty \right) h\gamma f(t-s)f(t+s) dt dh. \quad (20)$$

$$\text{Put } \left( 2 \int_{-s}^s - \int_{-\infty}^\infty \right) f(t-s)f(t+s) dt = G(s). \quad (21)$$

Then our postulate reduces to

$$\int_0^\infty s^\gamma G(s) ds = 0, \quad (22)$$

and we know from (10) that this is satisfied for all  $x_1, x_2$  if  $\gamma = 0$ . A sufficient condition for the absence of any other solution would be that  $G(s)$  shall change sign for precisely one value of  $s$ ; for if this value is  $s_0$ , we shall have

$$\int_0^\infty s_0^\gamma G(s) ds = 0, \quad (23)$$

and for positive  $\gamma$  the integrand in (22) is numerically larger than in (23) when  $s > s_0$  and smaller when  $s < s_0$ . Hence (22) cannot hold for any positive  $\gamma$ , and similarly for any negative  $\gamma$ . It has not been proved that  $G(s)$  has this property in general, but it has been verified for the cases where  $f(z) \propto \exp(-\frac{1}{2}z^2)$ ;  $f(z) = -\frac{1}{2}\exp\{-|z|\}$ ;  $f(z) = \frac{1}{2}$  for  $-1 < z < 1$ , and otherwise  $= 0$ ; and for a remarkable case suggested to me by Dr. A. C. Offord, where

$$f(z) = 1/2z^2 \quad (|z| > 1), \quad f(z) = 0 \quad (|z| < 1).$$

The property has an interesting analogue in the direct problem. Starting from (3) and putting  $x_1 + x_2 = 2a$ ,  $x_2 - x_1 = 2b$ , we find

$$P(da | b\alpha hH) = \frac{h^2 f\{h(a-b-\alpha)\}f\{h(a+b-\alpha)\}da}{\int_{-\infty}^{\infty} h^2 f\{h(a-b-\alpha)\}f\{h(a+b-\alpha)\}da}. \quad (24)$$

The condition that  $x_1 - \alpha$  and  $x_2 - \alpha$  shall have opposite signs is that  $|a - \alpha| < b$ . Hence for any  $b$  we can find the difference between the

chances that two observations with separation  $2b$  will have opposite signs or the same sign, and it is a positive multiple of

$$\left(2 \int_{-b}^b - \int_{-\infty}^{\infty}\right) hf\{h(a-b-\alpha)\}f\{h(a+b-\alpha)\} d(a-\alpha) = G(hb). \quad (25)$$

The fact that the integral of  $G(hb)$  over all values of  $b$  is zero means simply that the probabilities, given the law, that the first two observations will be on the same or opposite sides of the median are equal. For large  $b$  there will be an excess chance that they will be on opposite sides, for small  $b$  on the same side, and for continuous  $f(z)$  there will be a  $b$  such that the chances are equal. The result required is that there is only one such  $b$ ; this appears highly plausible but, as stated, has not been definitely proved except for special, though extremely different, forms of  $f(z)$ .

In a former presentation of the problem I took as a postulate that if  $x_1$  and  $x_2$  are the first two observations,  $x_2$  being the larger, and if  $x$  and  $\sigma$  are initially unknown, then

$$P(x_1 < x_3 < x_2 | x_1, x_2, H) = \frac{1}{3}, \quad (26)$$

I showed that only the rule

$$P(dxd\sigma | H) \propto dxd\sigma/\sigma \quad (27)$$

can lead to this. The former, however, was really derived from

$$P(x_1 < x_3 < x_2 | x, \sigma, H) = \frac{1}{3} \quad (28)$$

by an unconscious use of an argument analogous to that of 7.5. It is reasonable to say that the probability in (26) must be a constant independent of  $x_1$  and  $x_2$ , but with a different power of  $\sigma$  in (27) it would still be a constant but not  $\frac{1}{3}$ , and without some other principle it cannot be used to show that the  $d\sigma/\sigma$  rule is the only suitable one. The arguments for this are those of 3.1. But the principle (27) can be considered established otherwise, for complete previous ignorance of  $x$  and  $\sigma$ , and then we may ask whether we should expect it to be seriously altered if there is any vague information about  $\sigma$  such as we considered on p. 105. If the proper procedure is simply to truncate the prior probability law, and  $x_2 - x_1$  is much larger than the lower limit for  $\sigma$  and much smaller than the upper, the effect on the posterior probabilities will be negligible. This is in accordance with common sense. But if we used  $d\sigma/\sigma^{1+\gamma}$  we should be led to the result  $\frac{1}{2}$  for  $\gamma = 1$  and 0 for  $\gamma = -1$ . The latter is the uniform distribution for  $\sigma$ , and would lead also to the  $dx/|x|$  rule for the posterior probability of  $x$  from two

observations. Either would make a change in the probability distribution for  $x_3$  that cannot be accepted. We cannot admit that vague information about the range of possible values can make appreciable changes when the difference of the first two observations does not lie near either extreme, and we avoid this by simply truncating the law; and then we find that this makes a negligible difference to the result. The conclusion then is that vague information may as well be neglected and treated as total ignorance.

**3.8. Correlation.** Let the joint chance of two variables  $x$  and  $y$  be distributed according to the law

$$P(dxdy | \sigma, \tau, \rho, H) = \frac{1}{2\pi\sigma\tau(1-\rho^2)^{1/2}} \exp\left\{\frac{-1}{2(1-\rho^2)}\left(\frac{x^2}{\sigma^2} + \frac{y^2}{\tau^2} - \frac{2\rho xy}{\sigma\tau}\right)\right\} dxdy. \quad (1)$$

Then the joint chance of  $n$  pairs  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  is

$$P(\theta | \sigma, \tau, \rho, H) = \frac{1}{(2\pi\sigma\tau)^n(1-\rho^2)^{1/2n}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{Sx^2}{\sigma^2} + \frac{Sy^2}{\tau^2} - \frac{2\rho Sxy}{\sigma\tau}\right)\right\} dx_1 dy_1 \dots dx_n dy_n. \quad (2)$$

Put  $Sx^2 = ns^2$ ,  $Sy^2 = nt^2$ ,  $Sxy = nrst$ . Then  $s$ ,  $t$ , and  $r$  are sufficient statistics for  $\sigma$ ,  $\tau$ , and  $\rho$ .

We take  $\sigma$  and  $\tau$  as initially unknown. In accordance with what appears to be the natural interpretation of the correlation coefficient,  $\frac{1}{2}(1+\rho)$  may be regarded as a sampling ratio, being the ratio of the number of components that contribute to  $x$  and  $y$  with the same sign to the whole number of components. Thus the prior probability of  $\rho$ , in the most elementary case, can be taken as uniformly distributed, and

$$P(d\sigma d\tau d\rho | H) \propto d\sigma d\tau d\rho / \sigma\tau. \quad (3)$$

If  $\rho$  is near  $+1$  or  $-1$ , we may expect the rule to fail, for reasons similar to those given for sampling. But then it will usually happen also that one component contributes most of the variation, and the validity of the normal correlation surface itself will fail. The best treatment will then be to use the method of least squares. But in the typical case where the methods of correlation would be used we may adopt (3). Then, combining (2) with (3), we have

$$P(d\sigma d\tau d\rho | \theta H) \propto \frac{1}{(\sigma\tau)^n(1-\rho^2)^{1/2n}} \exp\left\{\frac{-n}{2(1-\rho^2)}\left(\frac{s^2}{\sigma^2} + \frac{t^2}{\tau^2} - \frac{2\rho rst}{\sigma\tau}\right)\right\} \frac{d\sigma d\tau d\rho}{\sigma\tau}. \quad (4)$$

The posterior probability distribution for  $\rho$  can be obtained by the substitution, due to Fisher,

$$\frac{st}{\sigma\tau} = \alpha, \quad \frac{s\tau}{\sigma t} = e^\beta; \quad (5)$$

whence 
$$\frac{s^2}{\sigma^2} = \alpha e^\beta, \quad \frac{t^2}{\tau^2} = \alpha e^{-\beta}, \quad \frac{\partial(\sigma, \tau)}{\partial(\alpha, \beta)} = \frac{st}{2\alpha^2}, \quad (6)$$

$$\begin{aligned} P(d\rho | \theta H) &\propto d\rho \int_0^\infty \int_{-\infty}^\infty \frac{\alpha^{n-1}}{(1-\rho^2)^{1/2n}} \exp\left\{-\frac{n\alpha}{1-\rho^2}(\cosh\beta - r\rho)\right\} d\alpha d\beta \\ &\propto d\rho \int_0^\infty \frac{(1-\rho^2)^{1/2n}}{(\cosh\beta - r\rho)^n} d\beta, \end{aligned} \quad (7)$$

since the integrand is an even function of  $\beta$ . At this stage the only function of the observations that is involved is  $r$ , so that  $r$  is a sufficient statistic for  $\rho$ . If we now put

$$\cosh\beta - r\rho = \frac{1-\rho r}{1-u} \quad (8)$$

the integral is transformed into

$$\frac{(1-\rho^2)^{1/2n}}{(1-\rho r)^{n-1/2}} \int_0^1 \frac{(1-u)^{n-1}}{\sqrt{(2u)}} \{1 - \frac{1}{2}(1+r\rho)u\}^{-1/2} du. \quad (9)$$

Since  $r$  and  $\rho$  are at most equal to 1, we can expand the last factor in powers of  $u$ , and integrate term by term, the coefficients being beta functions. Then, apart from an irrelevant factor, we find

$$P(d\rho | \theta H) \propto \frac{(1-\rho^2)^{1/2n}}{(1-\rho r)^{n-1/2}} S_n(\rho r) d\rho, \quad (10)$$

where 
$$S_n(\rho r) = 1 + \frac{1}{n+\frac{1}{2}} \frac{1+r\rho}{8} + \frac{1^2 \cdot 3^2}{2!(n+\frac{1}{2})(n+\frac{3}{2})} \left(\frac{1+r\rho}{8}\right)^2 + \dots, \quad (11)$$

a hypergeometric series. In actual cases  $n$  is usually large, and there is no appreciable error in reducing the series to its first term. But the form (10) is very asymmetrical. We see that the density is greatest near  $\rho = r$ , but since  $\rho$  must be between  $\pm 1$  there must be great asymmetry if  $r$  is not zero. This asymmetry can be greatly reduced by a transformation, also due to Fisher,

$$\tanh \zeta = \rho; \quad \tanh z = r; \quad \zeta = z + x, \quad (12)$$

so that the possible values of  $\zeta$  and  $z$  range between  $\pm\infty$ . This gives

$$\begin{aligned} P(d\zeta | \theta H) &\propto \frac{d\zeta}{\cosh^{n+2}\zeta \cosh^{n-1/2}z (1 - \tanh z \tanh \zeta)^{n-1/2}} \\ &\propto \frac{d\zeta}{\cosh^{5/2}\zeta \cosh^{-5/2}z \cosh^{n-1/2}x}, \end{aligned} \quad (13)$$

a power of  $\cosh z$  having been introduced to make the ordinate 1 at  $x = 0$ . The ordinate is a maximum where

$$-\frac{d}{dx} \left[ \frac{5}{2} \log \cosh \zeta + (n - \frac{1}{2}) \log \cosh x \right] = 0, \quad (14)$$

$$\text{or} \quad -\frac{5}{2} \tanh \zeta - (n - \frac{1}{2}) \tanh x = 0. \quad (15)$$

When  $n$  is large,  $x$  is small, and we have, nearly,

$$x = -\frac{5r}{2n}. \quad (16)$$

The second derivative is

$$-\frac{5}{2} \operatorname{sech}^2 \zeta - (n - \frac{1}{2}) \operatorname{sech}^2 x = -n \quad \text{nearly.} \quad (17)$$

$\operatorname{sech} \zeta$  can range from 0 to 1, so that the second derivative can range from  $-(n - \frac{1}{2})$  to  $-(n + 2)$ . Hence for large  $n$  we can write

$$\zeta = z - \frac{5r}{2n} \pm \frac{1}{\sqrt{n}}. \quad (18)$$

The distribution (13) is nearly symmetrical because the factor raised to a high power is  $\operatorname{sech} x$ , and it can be treated as nearly normal. Returning now to the series  $S_n(\rho r)$ , we see that its derivative with regard to  $\rho$  is of order  $1/n$ , and would displace the maximum ordinate by a quantity of order  $1/n^2$  if it was allowed for. But since the uncertainty is in any case about  $1/\sqrt{n}$  it is hardly worth while to allow for terms of order  $1/n$ , and those of order  $1/n^2$  can safely be omitted.

In most cases where the correlation coefficient arises, the distribution of chance is not centred on  $(0, 0)$  but on a pair of values  $(a, b)$ , which also have to be found from the observations. Then we must take

$$P(dadb d\sigma d\tau d\rho | H) \propto dadb d\sigma d\tau d\rho / \sigma\tau \quad (19)$$

and replace  $x$  and  $y$  in (1) by  $x-a$  and  $y-b$ . Then

$$\begin{aligned} &S \frac{(x-a)^2}{\sigma^2} + S \frac{(y-b)^2}{\tau^2} - 2\rho S \frac{(x-a)(y-b)}{\sigma\tau} \\ &= n \left\{ \frac{(a-\bar{x})^2}{\sigma^2} + \frac{(b-\bar{y})^2}{\tau^2} - \frac{2\rho(a-\bar{x})(b-\bar{y})}{\sigma\tau} \right\} + n \left( \frac{s^2}{\sigma^2} + \frac{t^2}{\tau^2} - \frac{2\rho rst}{\sigma\tau} \right), \end{aligned} \quad (20)$$

where now

$$\begin{aligned} n\bar{x} &= Sx, & n\bar{y} &= Sy, & ns^2 &= S(x-\bar{x})^2, & nt^2 &= S(y-\bar{y})^2, \\ nrst &= S(x-\bar{x})(y-\bar{y}). \end{aligned} \quad (21)$$

Then

$$\begin{aligned} P(dadb d\sigma d\tau d\rho | \theta H) &\propto \frac{1}{(\sigma\tau)^{n+1}(1-\rho^2)^{1/2n}} \exp \left[ \frac{-n}{2(1-\rho^2)} \left\{ \frac{(a-\bar{x})^2}{\sigma^2} + \frac{(b-\bar{y})^2}{\tau^2} - \frac{2\rho(a-\bar{x})(b-\bar{y})}{\sigma\tau} \right\} - \right. \\ &\quad \left. - \frac{n}{2(1-\rho^2)} \left\{ \frac{s^2}{\sigma^2} + \frac{t^2}{\tau^2} - \frac{2\rho rst}{\sigma\tau} \right\} \right] dadb d\sigma d\tau d\rho. \end{aligned} \quad (22)$$

Integration with regard to  $a$  and  $b$  then gives

$$P(d\sigma d\tau d\rho | \theta H) \propto \frac{1}{(\sigma\tau)^n(1-\rho^2)^{1/2(n-1)}} \exp \left\{ \frac{-n}{2(1-\rho^2)} \left( \frac{s^2}{\sigma^2} + \frac{t^2}{\tau^2} - \frac{2\rho rst}{\sigma\tau} \right) \right\} d\sigma d\tau d\rho. \quad (23)$$

Applying the transformations (5) and integrating with regard to  $\alpha$  and  $\beta$  will therefore only give an irrelevant function of  $n$  as a factor and replace  $n$  in (10) by  $n-1$ . Hence, in this case,

$$P(d\rho | \theta H) \propto \frac{(1-\rho^2)^{1/2(n-1)}}{(1-\rho r)^{n-3/2}} S_{n-1}(\rho r) d\rho \quad (24)$$

and, to the order retained,  $\zeta$  will still be given by (18). A slight change may perhaps be made with advantage in both cases. In the former, if  $n=1$ ,  $r$  will necessarily be  $\pm 1$  whatever  $\rho$  may be; in the latter this will hold for  $n=2$ . A permissible change will express this indeterminacy by making the uncertainty of  $\zeta$  infinite in these cases. Thus in the former we can write

$$\zeta = z - \frac{5r}{2n} \pm \frac{1}{\sqrt{(n-1)}}, \quad (25)$$

and in the latter 
$$\zeta = z - \frac{5r}{2n} \pm \frac{1}{\sqrt{(n-2)}}. \quad (26)$$

Fisher's theory of the correlation coefficient† follows different lines, but has suggested several points in the above analysis. He obtains the result that I should write

$$P(dr | a, b, \sigma, \tau, \rho, H) \propto \frac{(1-\rho^2)^{1/2(n-1)}(1-r^2)^{1/2(n-4)}}{(1-\rho r)^{n-3/2}} S_{n-1}(\rho r) dr, \quad (27)$$

and as this is independent of  $a, b, \sigma$ , and  $\tau$  we can drop these and replace the left side by  $P(dr | \rho H)$ . Also if we take the prior probability of  $\rho$

† *Biometrika*, **10**, 1915, 509-21; *Metron*, **1**, 1921, Part 4, 3-32.



as uniformly distributed, since  $r$  and  $dr$  are fixed for a given sample, this leads to

$$P(d\rho | rH) \propto \frac{(1-\rho^2)^{1/2(n-1)}}{(1-\rho r)^{n-3/2}} S_{n-1}(\rho r) d\rho, \quad (28)$$

which is identical with (24) except that the complete data  $\theta$  are replaced by  $r$ . This amounts to an alternative proof that  $r$  is a sufficient statistic for  $\rho$ ; the data contain no information relevant to  $\rho$  that is not contained in  $r$ .†

The bias shown by the second term in (25) and (26) is usually negligible, but requires attention if several equally correlated series are likely to be combined to give an improved estimate, since it always has the same sign. The question here will be, how far can the series be supposed mutually relevant? We cannot combine data from series with different correlation coefficients. But if the correlation is the same in all series we still have three cases.

1.  $a, b, \sigma, \tau$  the same in all series. Here the best method is to combine the data for all the series and find a summary value for  $r$  from them. The second term in  $\zeta$  will now be  $-5r/2 \sum n$ , which will be utterly negligible.

2.  $a, b$  different in the series,  $\sigma, \tau$  the same. Each pair  $(a, b)$  must now be eliminated separately and we shall be left with

$$P(d\sigma d\tau d\rho | \theta H) \propto \frac{1}{(\sigma\tau)^{\sum(n-1)+1} (1-\rho^2)^{1/2 \sum(n-1)}} \times \\ \times \exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{\sum ns^2}{\sigma^2} + \frac{\sum nt^2}{\tau^2} - \frac{2\rho \sum nrst}{\sigma\tau}\right)\right\} d\sigma d\tau d\rho. \quad (29)$$

The data, therefore, yield a summary correlation coefficient

$$R = \frac{\sum nrst}{(\sum ns^2)^{1/2} (\sum nt^2)^{1/2}} \quad (30)$$

and we proceed as before; the second term will be  $-5R/2 \sum (n-1)$ .

3.  $a, b, \sigma, \tau$  all different. Here  $\sigma$  and  $\tau$  must be eliminated for each series separately, before we can proceed to  $\rho$ . In this case we shall be led to the forms

$$P(d\rho | \theta H) \propto \frac{(1-\rho^2)^{1/2 \sum(n-1)}}{\prod (1-\rho r)^{(n-3/2)}} d\rho, \quad (31)$$

$$P(d\zeta | \theta H) \propto \frac{d\zeta}{\cosh^{1/2 p + 2} \zeta \prod \cosh^{n-3/2}(\zeta - z)}, \quad (32)$$

† *Proc. Roy. Soc. A*, **167**, 1938, 464-75.

where  $p$  is the number of series. The solution will therefore be, approximately,

$$\sum (n - \frac{1}{2})\zeta = \sum (n - \frac{1}{2})z - (\frac{1}{2}p + 2)\tanh \zeta \quad (33)$$

or, if we take  $Z$  as the weighted mean of the values of  $z$  and  $\tanh Z = R$ ,

$$\zeta = Z - \frac{\frac{1}{2}p + 2}{\sum n} R \pm \frac{1}{\sqrt{\{\sum (n - 2)\}}}. \quad (34)$$

The accuracy is similar to that of (18). The bias shown by the second term will in this case persist, and must be taken into account if many series are combined, since it will remain of the same order of magnitude while the standard error diminishes. This point is noticed by Fisher. The 2 in the numerator comes from the fact that if  $P(d\rho | H) \propto d\rho$ ,  $P(d\zeta | H) \propto \text{sech}^2 \zeta d\zeta$ . It therefore only appears once and its effect diminishes indefinitely as series are combined, but the extra  $\frac{1}{2}$  in (26) comes from the likelihood and is repeated in (34) by every series.

If  $\sigma$  and  $\tau$  in the correlation law are originally known, (4) will be replaced by

$$P(d\rho | \theta H) \propto \frac{1}{(1 - \rho^2)^{1/2n}} \exp \left\{ -\frac{n}{2(1 - \rho^2)} \left( \frac{s^2}{\sigma^2} + \frac{t^2}{\tau^2} - \frac{2\rho st}{\sigma\tau} \right) \right\} d\rho. \quad (35)$$

Thus  $r$  is no longer a sufficient statistic for  $\rho$ ;  $s$  and  $t$  are also relevant. The maximum posterior density is given by

$$\rho^3 - \rho^2 \frac{rst}{\sigma\tau} + \rho \left( \frac{s^2}{\sigma^2} + \frac{t^2}{\tau^2} - 1 \right) - \frac{rst}{\sigma\tau} = 0. \quad (36)$$

If  $r$  is positive, this is negative for  $\rho = 0$ , and equal to

$$\frac{s^2}{\sigma^2} + \frac{t^2}{\tau^2} - \frac{2rst}{\sigma\tau} \quad (37)$$

for  $\rho = +1$ , and this is positive. For  $\rho = r$  it is equal to

$$r \left( \frac{s^2}{\sigma^2} + \frac{t^2}{\tau^2} - 2 \right) + (r + r^3) \left( 1 - \frac{st}{\sigma\tau} \right), \quad (38)$$

which vanishes if  $s = \sigma$ ,  $t = \tau$ . Thus if  $s$  and  $t$  reach their expectations,  $r$  remains the best estimate of  $\rho$ . But (38) is negative if  $s/\sigma$  and  $t/\tau$  are very small, positive if they are large, and in the former case the best estimate of  $\rho$  will be larger, in the latter smaller than  $r$ . The reason is that if the scatters are unusually large it is evidence that too many large deviations have occurred; if there is a positive correlation at all and this is found in both variables, the most likely way for this to happen would be by way of an excess of deviations where  $x$  and  $y$  have the same sign, and the correlation in the sample would tend to be more than  $\rho$ .

It is unusual in practice, however, for  $\sigma$  and  $\tau$  to be well enough known for such supplementary information about  $\rho$  to be of much use.

**3.9. Invariance theory.** If we have two laws according to which the chances of a variable  $x$  being less than a given value are  $P$  and  $P'$ , any of the quantities

$$I_m = \int |(dP')^{1/m} - (dP)^{1/m}|^m, \quad J = \int \log \frac{dP'}{dP} d(P' - P) \quad (1)$$

has remarkable properties. They are supposed defined in the Stieltjes manner, by taking  $\delta P$ ,  $\delta P'$  for the same interval of  $x$ , forming the approximating sums, and then making the intervals of  $x$  tend to zero, and therefore may exist even if  $P$  and  $P'$  are discontinuous. They are all invariant for all non-singular transformations of  $x$  and of the parameters in the laws; and they are all positive definite. They can be extended immediately to joint distributions for several variables. They can therefore be regarded as providing measures of the discrepancy between two laws of chance. They are greatest if  $\delta P$  vanishes in all intervals where  $\delta P'$  varies and conversely; then  $I_m = 2$ ,  $J = \infty$ . They take these extreme values also if  $P$  varies continuously with  $x$ , and  $P'$  varies only at isolated values of  $x$ . The quantities  $I_2$  and  $J$  are specially interesting. Put  $p_r = \delta P_r$ ,  $p'_r = \delta P'_r$  for the interval  $\delta x_r$ . Let  $p_r$  depend on a set of parameters  $\alpha_i$  ( $i = 1$  to  $m$ ); and let  $p'_r$  be the result of changing  $\alpha_i$  to  $\alpha_i + \Delta\alpha_i$  where  $\Delta\alpha_i$  is small. Then, if  $p_r$  is differentiable with respect to  $\alpha_i$ , we have to the second order, using the summation convention with respect to  $i, k$ ,

$$J = \lim \sum_r \frac{1}{p_r} \left( \frac{\partial p_r}{\partial \alpha_i} \Delta\alpha_i \right) \left( \frac{\partial p_r}{\partial \alpha_k} \Delta\alpha_k \right) \quad (2)$$

$$= g_{ik} \Delta\alpha_i \Delta\alpha_k, \quad (3)$$

where

$$g_{ik} = \lim_{\Delta\alpha_i \rightarrow 0} \sum_r \frac{1}{p_r} \frac{\partial p_r}{\partial \alpha_i} \frac{\partial p_r}{\partial \alpha_k}. \quad (4)$$

Also

$$I_2 = \frac{1}{4} g_{ik} \Delta\alpha_i \Delta\alpha_k \quad (5)$$

to the same accuracy. Thus  $J$  and  $4I_2$  have the form of the square of an element of distance in curvilinear coordinates. If we transform to any other set of parameters  $\alpha'_j$ ,  $J$  and  $4I_2$  are unaltered, and

$$J = g'_{jl} \Delta\alpha'_j \Delta\alpha'_l, \quad (6)$$

where

$$g'_{jl} = g_{ik} \frac{\partial \alpha_i}{\partial \alpha'_j} \frac{\partial \alpha_k}{\partial \alpha'_l}. \quad (7)$$

Then

$$||g'_{jl}|| = ||g_{ik}|| \left| \left| \frac{\partial \alpha_i}{\partial \alpha'_j} \right| \right| \left| \left| \frac{\partial \alpha_k}{\partial \alpha'_i} \right| \right|. \quad (8)$$

But in the transformation of a multiple integral

$$d\alpha_1 d\alpha_2 \dots d\alpha_m = \left| \left| \frac{\partial \alpha_i}{\partial \alpha'_j} \right| \right| d\alpha'_1 \dots d\alpha'_m \quad (9)$$

$$= \left( \frac{||g'_{jl}||}{||g_{ik}||} \right)^{1/2} d\alpha'_1 \dots d\alpha'_m. \quad (10)$$

$$\text{Hence} \quad ||g_{ik}||^{1/2} d\alpha_1 \dots d\alpha_m = ||g'_{jl}||^{1/2} d\alpha'_1 \dots d\alpha'_m. \quad (11)$$

This expression is therefore invariant for all non-singular transformations of the parameters. It is not known whether any analogous forms can be derived from  $I_m$  if  $m \neq 2$ ; but the form of  $I_m$  is then usually much more complicated.

In consequence of this result, if we took the prior probability density for the parameters to be proportional to  $||g_{ik}||^{1/2}$ , it could be stated for any law that is differentiable with respect to all parameters in it, and would have the property that the total probability in any region of the  $\alpha_i$  would be equal to the total probability in the corresponding region of the  $\alpha'_i$ ; in other words, it satisfies the rule that equivalent propositions have the same probability. Consequently any arbitrariness in the choice of the parameters could make no difference to the results, and it is proved that for this wide class of laws a consistent theory of probability can be constructed. Hence our initial requirement 2 (p. 8) can be satisfied for this class; it remains to be seen whether the desirable, but less precise or fundamental requirement 7 (p. 10) is also satisfied.

For the normal law of error

$$p_r \doteq \frac{1}{\sqrt{(2\pi)\sigma}} \exp\left\{-\frac{(x_r - \lambda)^2}{2\sigma^2}\right\} \delta x_r, \quad (12)$$

we have exactly, if

$$\sigma = \sigma_0 e^{-1/2\zeta}, \quad \sigma' = \sigma_0 e^{1/2\zeta}, \quad (13)$$

$$\begin{aligned} I_2 &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{(2\pi)}} \left[ \frac{1}{\sqrt{\sigma'}} \exp\left\{-\frac{(x-\lambda')^2}{4\sigma'^2}\right\} - \frac{1}{\sqrt{\sigma}} \exp\left\{-\frac{(x-\lambda)^2}{4\sigma^2}\right\} \right]^2 dx \\ &= 2 \left[ 1 - \frac{\sqrt{2}}{\sqrt{(\sigma'/\sigma + \sigma/\sigma')}} \exp\left\{-\frac{(\lambda' - \lambda)^2}{4(\sigma^2 + \sigma'^2)}\right\} \right] \\ &= 2 \left[ 1 - \operatorname{sech}^{1/2} \zeta \exp\left\{-\frac{(\lambda' - \lambda)^2}{8\sigma_0^2 \cosh \zeta}\right\} \right]; \end{aligned} \quad (14)$$

$$\begin{aligned}
J &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{(2\pi)}} \left\{ -\log \frac{\sigma'}{\sigma} + \frac{(x-\lambda)^2}{2\sigma^2} - \frac{(x-\lambda')^2}{2\sigma'^2} \right\} \times \\
&\quad \times \left[ \frac{1}{\sigma'} \exp \left\{ -\frac{(x-\lambda')^2}{2\sigma'^2} \right\} - \frac{1}{\sigma} \exp \left\{ -\frac{(x-\lambda)^2}{2\sigma^2} \right\} \right] dx \\
&= \frac{1}{2} \left( \frac{\sigma'}{\sigma} - \frac{\sigma}{\sigma'} \right)^2 + \frac{1}{2} \left( \frac{1}{\sigma^2} + \frac{1}{\sigma'^2} \right) (\lambda' - \lambda)^2 \\
&= 2 \sinh^2 \zeta + \cosh \zeta \frac{(\lambda' - \lambda)^2}{\sigma^2}.
\end{aligned} \tag{15}$$

To the second order

$$4I_2 = J = 2 \left( \frac{d\sigma}{\sigma} \right)^2 + \left( \frac{d\lambda}{\sigma} \right)^2. \tag{16}$$

Three cases arise. If  $\sigma$  is fixed, the coefficient of  $(d\lambda)^2$  is constant, giving a uniform prior probability distribution for  $\lambda$  over the range permitted, in accordance with the rule for a location parameter. If  $\lambda$  is fixed,  $||g_{ik}||^{1/2} d\sigma \propto d\sigma/\sigma$ , again in accordance with the rule that we have adopted. This rule, of course, has itself been chosen largely for reasons of invariance under transformation of  $\sigma$ . But if  $\lambda$  and  $\sigma$  are both varied,  $||g_{ik}||^{1/2} d\lambda d\sigma \propto d\lambda d\sigma/\sigma^2$  instead of  $d\lambda d\sigma/\sigma$ . If the same method was applied to a joint distribution for several variables about independent true values, an extra factor  $1/\sigma$  would appear for each. The index in the corresponding  $t$  distribution would always be  $\frac{1}{2}(n+1)$ , however many true values were estimated. This is unacceptable. In the usual situation in an estimation problem  $\lambda$  and  $\sigma$  are each capable of any value over a considerable range, and neither gives any appreciable information about the other. Then if we are given  $-M < \lambda < M$ ,  $\sigma_1 < \sigma < \sigma_2$ , we should take

$$\begin{aligned}
P(d\lambda | H) &= d\lambda/2M, & P(d\sigma | H) &= d\sigma/\sigma \log(\sigma_2/\sigma_1), \\
P(d\lambda d\sigma | H) &= P(d\lambda | H)P(d\sigma | H) = \frac{d\lambda d\sigma}{2M\sigma \log(\sigma_2/\sigma_1)}.
\end{aligned} \tag{17}$$

The departure from the general rule is thus explicable as due to the use of a previous judgement of irrelevance.

There is no trouble for  $\sigma$  alone or  $\lambda$  alone; it arises when they are considered both at once. Now take a law such as that of partial correlation

$$P(dx_1 \dots dx_n | \alpha_{ik}, \sigma_i, H) = A \exp(-\frac{1}{2}W) \prod dx_r,$$

where

$$W = \sum \alpha_{ik} x_i x_k / \sigma_i \sigma_k$$

and the  $x_i$  are a set of observables. Here for each  $x_i$  there is a corresponding scale parameter  $\sigma_i$  and the  $\alpha_{ik}$  are numerical coefficients. It is

clear from considerations of similarity that  $J$ , to the second order, is a quadratic in  $(d\sigma_i/\sigma_i)$ , and that  $||g_{ik}||$  will be of the form  $\prod_i \sigma_i^{-2} B$ , where  $B$  is a numerical factor depending on the  $\alpha_{ik}$ . Hence the rule leads to

$$P(d\sigma_i d\alpha_{km} | H) \propto \prod_i (d\sigma_i/\sigma_i) B^{1/2} \prod d\alpha_{km}, \quad (18)$$

which is what we should expect. There is no difficulty in the introduction of any number of scale parameters.

We can then deal with location parameters, on the hypothesis that the scale and numerical parameters are irrelevant to them, by simply taking their prior probability uniform. If  $\lambda$  and  $\sigma$  are location and scale parameters in general, and the numerical parameters are  $\alpha_i$ , we can take

$$P(d\lambda d\sigma \prod d\alpha_i | H) \propto d\lambda ||g_{ik}||^{1/2} d\sigma \prod d\alpha_i, \quad (19)$$

where  $||g_{ik}||$  is found by varying only  $\sigma$  and the  $\alpha_i$ , and is equal to  $1/\sigma$  times a function of the  $\alpha_i$ . This is invariant for transformations of the form

$$\lambda' = \lambda + af(\alpha_i), \quad (20)$$

which is the only form of transformation of  $\lambda$  that we should wish to make.

If  $\sigma$  is already uniquely defined, a satisfactory rule would be

$$P(d\lambda d\sigma \prod d\alpha_i | H) \propto d\lambda \frac{d\sigma}{\sigma} ||g_{ik}||^{1/2} \prod d\alpha_i, \quad (21)$$

where  $g_{ik}$  is now found by varying only the  $\alpha_i$ , keeping  $\lambda, \sigma$  constant.

Again, take a Pearson Type I law  $A(x-c_1)^{m_1}(c_2-x)^{m_2}dx$ . For any non-zero change of  $c_1$  or  $c_2$ ,  $J$  is infinite.  $I_2$  is not of the second order in  $\Delta c_1, \Delta c_2$  unless  $m_1, m_2 \geq 1$ . If we evaluate the coefficients in the differential form by integration, e.g.

$$g_{c_1 c_1} = \int_{c_1}^{c_2} \frac{1}{A} \left( \frac{\partial A}{\partial c_1} - \frac{A m_1}{x - c_1} \right)^2 (x - c_1)^{m_1} (c_2 - x)^{m_2} dx. \quad (22)$$

This diverges unless  $m_1 > 1$ . Thus the general rule fails if the law is not differentiable at a terminus. But the case where either of  $m_1, m_2 < 1$  is precisely the case where a terminus can be estimated from  $n$  observations with an uncertainty  $o(n^{-1/2})$ , and it is then advantageous to take that terminus as a parameter explicitly; the occasion for transformation of it no longer exists. If one of  $m_1, m_2 \leq 1$  it is natural to take  $c_1$  or  $c_2$  respectively as location parameter; if both are  $\leq 1$ , it is equally natural to take  $\frac{1}{2}(c_1 + c_2)$  as location parameter and  $\frac{1}{2}(c_2 - c_1)$  as scale parameter. In either case we need only evaluate the differential form for changes of the other parameters and find a prior probability for them independent

of  $c_1$ ,  $c_2$ , or both, as the case may be. It is interesting to find that an apparent failure of the general rule corresponds to a well-known exceptional case in an estimation problem and that the properties of this case themselves suggest the appropriate modification of the procedure.

For the comparison of two chances  $\alpha$ ,  $\alpha'$  we have

$$\begin{aligned} I_2 &= (\sqrt{\alpha'} - \sqrt{\alpha})^2 + \{\sqrt{(1-\alpha')} - \sqrt{(1-\alpha)}\}^2 \\ &= 2 - 2\sqrt{(\alpha\alpha')} - 2\sqrt{(1-\alpha)(1-\alpha')}. \end{aligned} \quad (23)$$

This takes a simple form if we put  $\alpha = \sin^2 a$ ,  $\alpha' = \sin^2 a'$ ;

$$I_2 = 4 \sin^2 \frac{1}{2}(a' - a) \div (a' - a)^2. \quad (24)$$

The exact form of  $J$  is more complicated:

$$J = (\alpha' - \alpha) \log \frac{\alpha'(1-\alpha)}{\alpha(1-\alpha')}. \quad (25)$$

Then the rule (11) gives

$$P(d\alpha | H) = \frac{2}{\pi} d\alpha = \frac{1}{\pi} \frac{d\alpha}{\sqrt{\{\alpha(1-\alpha)\}}}. \quad (26)$$

This is an interesting form, because we have already had hints that both the usual rule  $d\alpha$  and Haldane's rule

$$P(d\alpha | H) \propto \frac{d\alpha}{\alpha(1-\alpha)}$$

are rather unsatisfactory, and that something intermediate would be better.

For a set of chances  $\alpha_r$  ( $r = 1, \dots, m$ ,  $\sum \alpha_r = 1$ ) we find

$$I_2 = 2 - 2 \sum \sqrt{\{\alpha_r(\alpha_r + \Delta\alpha_r)\}} \div \frac{1}{4} \sum \frac{(\Delta\alpha_r)^2}{\alpha_r} \quad (27)$$

$$= \frac{1}{4} \sum_{r=1}^{m-1} \frac{(\Delta\alpha_r)^2}{\alpha_r} + \frac{1}{4} \frac{\left(\sum_{r=1}^{m-1} \Delta\alpha_r\right)^2}{\alpha_m}.$$

Then

$$||g_{ik}|| = \frac{1}{\prod \alpha_r}, \quad (28)$$

$$P(d\alpha_1 \dots d\alpha_{m-1} | H) \propto \frac{d\alpha_1 \dots d\alpha_{m-1}}{\sqrt{\left(\prod_1^m \alpha_r\right)}}. \quad (29)$$

The rule so found is an appreciable modification of the rule for multiple sampling given in 3.23, and is the natural extension of (26).

If  $\phi_r$ ,  $\psi_s$  are two sets of exhaustive and exclusive alternatives,  $\phi_r$  being irrelevant to  $\psi_s$ , with chances  $\alpha_r$ ,  $\beta_s$  ( $r = 1$  to  $m$ ,  $s = 1$  to  $n$ ) the chance

of  $\phi_r, \psi_s$  is  $\alpha_r \beta_s$ . If we vary both  $\alpha_r$  and  $\beta_s$  and consider  $I_2$  and  $J$  for the changes of  $\alpha_r \beta_s$ , we get

$$\begin{aligned} I_2 &= 2 - 2 \sum \sum \sqrt{(\alpha_r \beta_s \alpha'_r \beta'_s)} \\ &= 2 - 2(1 - \tfrac{1}{2} I_{2,\alpha})(1 - \tfrac{1}{2} I_{2,\beta}), \end{aligned} \quad (30)$$

$$\begin{aligned} J &= \sum \sum (\alpha'_r \beta'_s - \alpha_r \beta_s) \log \frac{\alpha'_r \beta'_s}{\alpha_r \beta_s} \\ &= \sum (\alpha'_r - \alpha_r) \log \frac{\alpha'_r}{\alpha_r} + \sum (\beta'_s - \beta_s) \log \frac{\beta'_s}{\beta_s} \\ &= J_\alpha + J_\beta, \end{aligned} \quad (31)$$

suffixes  $\alpha, \beta$  indicating the values if  $\alpha_r, \beta_s$  are varied separately. Hence for probabilities expressible as products of chances  $\log(1 - \tfrac{1}{2} I_2)$  and  $J$  have an exact additive property. The estimation rule then gives

$$P(d\alpha_1 \dots d\alpha_{m-1} d\beta_1 \dots d\beta_{n-1} | H) = P(d\alpha_1 \dots d\alpha_{m-1} | H) P(d\beta_1, \dots, d\beta_{n-1} | H), \quad (32)$$

which is satisfactory.

Now consider a set of quantitative laws  $\phi_r$  with chances  $\alpha_r$ . If  $\phi_r$  is true, the chance of a variable  $x$  being in a range  $dx$  is  $f_r(x, \alpha_{r1}, \dots, \alpha_{rn}) dx$ , and

$$P(\phi_r dx | \alpha_r, \alpha_{rs}, H) = \alpha_r f_r(x, \alpha_{r1}, \dots, \alpha_{rn}) dx. \quad (33)$$

For variations of both the  $\alpha_r$  and the  $\alpha_{rs}$ ,

$$\begin{aligned} I_2 &= 2 - 2 \sum \sqrt{\{\alpha_r(\alpha_r + \Delta\alpha_r)\}} \int \sqrt{\{f_r(f_r + \Delta f_r)\}} dx \\ &= 2 - \sum \sqrt{\{\alpha_r(\alpha_r + \Delta\alpha_r)\}} (2 - I_{2,r}) \\ &= I_{2,\alpha} + \sum \sqrt{\{\alpha_r(\alpha_r + \Delta\alpha_r)\}} I_{2,r} \end{aligned} \quad (34)$$

and, to the second order,

$$I_2 = I_{2,\alpha} + \sum \alpha_r I_{2,r}. \quad (35)$$

$I_{2,r}$  is the discrepancy between  $f_r$  with parameters  $\alpha_{rs}$  and with parameters  $\alpha_{rs} + \Delta\alpha_{rs}$ . If we form  $\|g_{ik}\|^{1/2}$  for variations of all  $\alpha_r$  and all  $\alpha_{rs}$ , the rule will then give the same factor depending on the  $\alpha_{rs}$  as for estimation of  $\alpha_{rs}$  when  $\phi_r$  is taken as certain. But for every  $\alpha_{rs}$  a factor  $\alpha_r^{1/2}$  will enter into  $\|g_{ik}\|^{1/2}$ , and will persist on integration with regard to the  $\alpha_{rs}$ . Hence the use of the rule for all  $\alpha_r$  and all  $\alpha_{rs}$  simultaneously would lead to a change of the prior probability of  $\alpha_r$  for every parameter contained in  $f_r$ . This would not be inconsistent, but as for scale parameters it is not the usual practical case.  $\alpha_r$  is ordinarily determined only by the conditions of sampling and has nothing to do with the complexity of the  $f_r$ . To express this, we need a modification analogous to that used for location parameters; the chance  $\alpha_r$ , like a location parameter, must be put in a privileged position, and we have to consider what type of invariance can hold for it.



The general form (11) gives invariance for the most general non-singular transformations of the parameters. In this problem it would permit the use of a set of parameters that might be any independent functions of both the  $\alpha_r$  and the  $\alpha_{rs}$ . In sampling for discrete alternatives it is not obvious that there is any need to consider transformations of the chances at all.

If we take

$$P(\prod d\alpha_r \prod d\alpha_{rs} | H) \propto \frac{\prod_1^{m-1} d\alpha_r}{\sqrt{(\prod_1^m \alpha_r)}} \prod_{r=1}^m \|g_{ik}\|_r^{1/2} \prod_{s=1}^{n_r} d\alpha_{rs}, \quad (36)$$

where  $\|g_{ik}\|_r$  is based on comparison of  $f_r$  with  $f_r + \Delta f_r$ , we shall still have invariance for all transformations of the  $\alpha_r$  among themselves and of the  $\alpha_{rs}$  among themselves, and this is adequate. If we do not require to consider transformations of the  $\alpha_r$  we do not need the factor  $(\prod \alpha_r)^{-1/2}$ . If some of the  $\alpha_{rs}$  are location and scale parameters, we can use the modification (19). (36) can then be regarded as the appropriate extension of (32), which represents the case where  $\alpha_{rs} \approx \beta_s$ , independent of  $r$ .

For the Poisson law

$$P(m | rH) = e^{-r} \frac{r^m}{m!} \quad (37)$$

we find

$$\left. \begin{aligned} I_2 &= 2 - 2 \exp\{-\tfrac{1}{2}(\sqrt{r'} - \sqrt{r})^2\}, \\ J &= (r' - r) \log(r'/r), \end{aligned} \right\} \quad (38)$$

leading to

$$P(dr | H) \propto dr/\sqrt{r}. \quad (39)$$

This conflicts with the rule  $dr/r$  used in 3.3, which was quite satisfactory. The Poisson parameter, however, is in rather a special position. It is usually the product of a scale factor with an arbitrary sample size, which is not chosen until we already have some information about the probable range of values of the scale parameter. It does, however, point a warning for all designed experiments. The whole point of general rules for the prior probability is to give a starting-point, which we take to represent previous ignorance. They will not be correct if previous knowledge is being used, whether it is explicitly stated or not. In the case of the Poisson law the sample size is chosen so that  $r$  will be a moderate number, usually 1 to 10; we should not take it so that the chance of the event happening at all is very small. The  $dr/r$  rule, in fact, may express complete ignorance of the scale parameter; but  $dr/\sqrt{r}$  may express just enough information to suggest that the experiment is worth making. Even if we used (39), the posterior probability density after one observation would be integrable over all  $r$ .

For normal correlation we get

$$J = -2 + \frac{\sigma^2/\sigma'^2 + \tau^2/\tau'^2 - 2\rho\rho'\sigma\tau/\sigma'\tau'}{2(1-\rho'^2)} + \frac{\sigma'^2/\sigma^2 + \tau'^2/\tau^2 - 2\rho\rho'\sigma'\tau'/\sigma\tau}{2(1-\rho^2)}, \quad (40)$$

$$I_2 = 2 - 4(\sigma\sigma'\tau\tau')^{1/2}(1-\rho^2)^{1/4}(1-\rho'^2)^{1/4} \times \\ \times \{\sigma'^2\tau'^2(1-\rho'^2) + \sigma^2\tau^2(1-\rho^2) + \sigma'^2\tau^2 + \sigma^2\tau'^2 - 2\rho\rho'\sigma\sigma'\tau\tau'\}^{-1/2}. \quad (41)$$

If we put

$$\sigma' = \sigma e^{2u}, \quad \tau' = \tau e^{2v}, \quad \rho = \tanh \zeta, \quad \rho' = \tanh \zeta' \quad (42)$$

and change the parameters to  $\zeta, u+v, u-v$ , we get, to the second order in  $u, v, \zeta' - \zeta$ ,

$$J = (1 + \tanh^2 \zeta)(\zeta' - \zeta)^2 - \\ - 4 \tanh \zeta(\zeta' - \zeta)(u+v) + 4(u+v)^2 + 4(u-v)^2 \cosh^2 \zeta, \quad (43)$$

$$||g_{ik}|| = 64 \cosh^2 \zeta, \quad (44)$$

$$P(d\sigma d\tau d\rho | H) \propto \frac{d\sigma d\tau}{\sigma\tau} \frac{d\rho}{(1-\rho^2)^{3/2}}. \quad (45)$$

The modifications of the analysis of 3.8, when this rule is adopted, are straightforward. The divergence at  $\rho = \pm 1$  is a new feature, and persists if there is one observation, when  $\tau$  is  $\pm 1$ . If there are two observations and  $\tau \neq \pm 1$  the posterior probability density for  $\rho$  has a convergent integral, so that the rule gives intelligible answers when the data have anything useful to say.

In problems concerned with correlations the results will depend somewhat on the choice of parameters in defining  $J$ . From (43) we can write  $J$  for small variations as

$$J = (\zeta' - \zeta)^2 + 4 \cosh^2 \zeta (u-v)^2 + \{2(u+v) - \tanh \zeta(\zeta' - \zeta)\}^2. \quad (46)$$

Now  $\sigma$  and  $\tau$  can be regarded as parameters defined irrespectively of  $\rho$ ; for whatever  $\rho$  may be, the probability distributions of  $x$  and  $y$  separately are normal with standard errors  $\sigma, \tau$ . Thus we may analyse the estimation of a correlation into three parts: what is the probability distribution of  $x$ ? what is that of  $y$ ? and given those of  $x$  and  $y$  separately, does the variation of  $y$  depend on that of  $x$ , and conversely? In this analysis we are restricted to a particular order of testing and in giving the prior probability of  $\zeta$  we should evaluate  $J$  with  $\sigma$  and  $\tau$  fixed. In this case (40) becomes

$$J = \frac{(1+\rho\rho')(\rho-\rho')^2}{(1-\rho^2)(1-\rho'^2)} \quad (47)$$

and

$$P(d\rho | \sigma\tau H) \propto \frac{(1+\rho^2)^{1/2}}{1-\rho^2} d\rho. \quad (48)$$

From the interpretation of a correlation coefficient in terms of a chance (2.5) we should have expected

$$P(d\rho | \sigma\tau H) = \frac{1}{\pi} \frac{d\rho}{\sqrt{(1-\rho^2)}}. \quad (49)$$

This is integrable as it stands and would be free from objection in any case where the model considered in 2.3 is known to be representative of the physics of the problem.

The different rules for  $\rho$  correspond to rather different requirements. (45) contemplates transformations of  $\rho$ ,  $\sigma$ ,  $\tau$  together, (48) transformations only of  $\rho$ , keeping  $\sigma$ ,  $\tau$  fixed. (49) does not contemplate transformations at all, but appeals to a model. But the rule for this model itself is derived by considering transformations of a simple chance, and the need for this is not obvious. We really cannot say that any of these rules is better than the uniform distribution adopted in 3.8.

These rules do not cover the sampling of a finite population. The possible numbers of one type are then all integers and differentiation is impossible. This difficulty does not appear insuperable. Suppose that the population is of number  $n$  and contains  $r$  members with the property. Treat this as a sample of  $n$  derived from a chance  $\alpha$ . Then

$$\begin{aligned} P(d\alpha | nH) &= \frac{d\alpha}{\pi\sqrt{\{\alpha(1-\alpha)\}}}, \\ P(r | n, \alpha H) &= \frac{n!}{r!(n-r)!} \alpha^r (1-\alpha)^{n-r}, \\ P(r d\alpha | nH) &= \frac{n!}{\pi r! (n-r)!} \alpha^{r-1/2} (1-\alpha)^{n-r-1/2} d\alpha, \\ P(r | nH) &= \frac{(r-\frac{1}{2})! (n-r-\frac{1}{2})!}{\pi r! (n-r)!}. \end{aligned} \quad (50)$$

This is finite both for  $r = 0$  and  $r = n$ .

To sum up the results found so far:

1. A widely applicable rule is available for assessing the prior probability in estimation problems and will satisfy the requirement of consistency whenever it can be applied, in the sense that it is applicable under any non-singular transformation of the parameters, and will lead to equivalent results. At least this proves the possibility of a consistent theory of induction, covering a large part of the subject.

2. There are many cases where the rule, though consistent, leads to results that appear to differ too far from current practice, but it is still possible to use modified forms of the rule which actually have a wider

applicability. These cases are associated with conditions where there is reason to take the prior probabilities of some of the parameters as independent of one another.

3. The rule is not applicable to laws that are not differentiable with regard to all parameters in them; but in this case a modification of the rule is often satisfactory.

4. In some cases where the parameters themselves can take only discrete values, an extension of the rule is possible.

Further investigation is desirable; there may be some other method that would preserve or even extend the generality of the one just discussed, while dealing with some of the awkward cases more directly.

## IV

### APPROXIMATE METHODS AND SIMPLIFICATIONS

'Troll, to thyself be true—enough.'

IBSEN, *Peer Gynt*.

**4.0. Maximum likelihood.** If a law containing parameters  $\alpha, \beta, \gamma, \dots$  and a set of observations  $\theta$  lead to the likelihood function  $L(\alpha, \beta, \gamma, \dots)$ , and if the prior probability is

$$P(d\alpha d\beta d\gamma \dots | H) \propto f(\alpha, \beta, \gamma, \dots) d\alpha d\beta d\gamma \dots, \quad (1)$$

$$\text{then } P(d\alpha d\beta d\gamma \dots | \theta H) \propto f(\alpha, \beta, \gamma, \dots) L(\alpha, \beta, \gamma, \dots) d\alpha d\beta d\gamma \dots. \quad (2)$$

There will in general be a set of values of  $\alpha, \beta, \gamma, \dots$ , say  $a, b, c, \dots$  that make  $L$  a maximum. These may be called the 'maximum likelihood solution'. Then if we put  $\alpha = a + \alpha'$ , and so on, we can usually expand  $\log f$  and  $\log L$  in powers of  $\alpha', \beta', \gamma', \dots$ . Now the maximum posterior probability density is given by

$$\frac{1}{L} \frac{\partial L}{\partial \alpha} + \frac{1}{f} \frac{\partial f}{\partial \alpha} = 0 \quad (3)$$

with similar equations. The prior probability function  $f$  is independent of  $n$ , the number of observations;  $\log L$  in general increases like  $n$ . Hence if  $(\alpha', \beta', \gamma', \dots)$  satisfy (3), they will be of order  $1/n$ .

Also, if we neglect terms of order above the second in  $\log L$  and  $\log f$ , the second derivatives of  $\log Lf$  will contain terms of order  $n$  from  $\log L$ , while those from  $\log f$  do not increase. Hence for  $\alpha', \beta', \gamma', \dots$  small, the quadratic terms will be

$$-n\phi_2(\alpha', \beta', \gamma', \dots) + O(\alpha'^2, \beta'^2, \gamma'^2, \dots), \quad (4)$$

where  $\phi_2$  is a positive quadratic form independent of  $n$ . Hence the posterior probability is concentrated in ranges of order  $n^{-1/2}$ , and this indicates the uncertainty of any possible estimates of  $\alpha, \beta, \gamma, \dots$ . But the differences between the values that make the likelihood and the posterior density maxima are only of order  $1/n$ . Hence if the number of observations is large, the error committed by taking the maximum likelihood solution as the estimate is less than the uncertainty inevitable in any case. Further, the terms in  $\log Lf$  that come from  $L$  are of order  $n$  times those from  $f$ , and hence if we simply take the posterior density proportional to  $L$  we shall get the right uncertainties within factors of order  $1/n$ . Thus the errors introduced by treating the prior probability as uniform will be of no practical importance if the number of observations is large.

The method of maximum likelihood has been vigorously advocated

by Fisher; the above argument shows that in the great bulk of cases its results are indistinguishable from those given by the principle of inverse probability, which supplies a justification of it. An accurate statement of the prior probability is not necessary in a pure problem of estimation when the number of observations is large. What the result amounts to is that unless we previously know so much about the parameters that the observations can tell us little more, we may as well use the prior probability distribution that expresses ignorance of their values; and in cases where this distribution is not yet known there is no harm in taking a uniform distribution for any parameter that cannot be infinite. The difference made by any ordinary change of the prior probability is comparable with the effect of one extra observation.

Even where the uncertainty is of order  $1/n$  instead of  $1/n^{1/2}$  this may still be true. Thus for the rectangular distribution we had  $L \propto \sigma^{-n}$ , while  $Lf \propto \sigma^{-n-1}$ . The differences between the ranges for a given probability that the quantity lies within them, obtained by using  $L$  instead of  $Lf$ , will be of order  $1/n$  of the ranges themselves.

**4.01. Relation of maximum likelihood to invariance theory.** Another important consequence of (1) and (3) is as follows. In 4.0 (2) we have, taking the case of three unknowns,

$$P(\theta | \alpha\beta\gamma H) \propto L,$$

where  $L$  depends on the observations and on  $\alpha, \beta, \gamma$ .  $a, b, c$  are the values of  $\alpha, \beta, \gamma$  that make  $L$  a maximum, the observations being kept the same. Then for given  $\alpha, \beta, \gamma$  we can find by integration a probability that  $a, b, c$  lie in given intervals  $da, db, dc$ . This does not assume that  $a, b, c$  are sufficient statistics. Then when  $n$  is large  $L$  is nearly proportional to

$$\exp\{-\frac{1}{2}ng_{ik}(\alpha_i - a_i)(\alpha_k - a_k)\} \prod da_m$$

and all parameters given by maximum likelihood tend to become sufficient statistics. Further, the constant factor is  $(n/2\pi)^{1/2m} ||g_{ik}||^{1/2}$ , and it is of trivial importance whether  $g_{ik}$  is evaluated for the actual values  $\alpha_i$  or for  $\alpha_i = a_i$ . Hence if we use  $||g_{ik}||^{1/2}$  for the prior probability density, the probability distribution of  $\alpha_i - a_i$  is nearly the same when  $n$  is large, whether it is taken on data  $\alpha_i$  or  $a_i$ ; this is irrespective of the actual value of  $\alpha_i$ .

Mr. P. H. Diananda has suggested, on this account, that we could state an invariance rule for the prior probability in estimation problems as follows. Take, for  $n$  large,

$$P(\alpha_i < a_i < \alpha_i + d\alpha_i | \alpha_i H) = f(\alpha_i) \prod d\alpha_i,$$

where  $i$  covers all parameters in the law; then if we take

$$P(d\alpha_i | H) \propto f(\alpha_i) \prod d\alpha_i,$$

we have a rule equivalent to the  $||g_{ik}||^{1/2}$  rule where the latter is applicable. It also works for the rectangular distribution. A similar rule was given independently by Mr. Wilfred Perks, who, however, considered only one parameter.†

Again, in the argument of 3.9 we considered only the values of the invariants for one observation, except that we showed that for sets of observations derived independently from the laws  $J$  and  $\log(1 - \frac{1}{2}I_2)$  have an additive property. This argument is no longer applicable if the observations are not derived independently; this happens in problems where the law predicts something about the order of occurrence as well as about their actual values. But it now appears that we can consistently extend the rule to cover such cases. If two laws give

$$P(\theta | \alpha_i H) = L(\theta, \alpha_i); \quad P(\theta | \alpha'_i H) = L(\theta, \alpha'_i),$$

we can take

$$J = \lim_{n \rightarrow \infty} \frac{1}{n} \sum \log \frac{L(\theta, \alpha'_i)}{L(\theta, \alpha_i)} \{L(\theta, \alpha'_i) - L(\theta, \alpha_i)\},$$

$$-\log(1 - \frac{1}{2}I_2) = -\lim_{n \rightarrow \infty} \frac{1}{n} \sum \log[1 - \frac{1}{2}\{L^{1/2}(\theta, \alpha'_i) - L^{1/2}(\theta, \alpha_i)\}^2],$$

summations being over the possible values of  $\theta$ . Both reduce correctly when the observations are derived independently.

**4.1. An approximation to maximum likelihood.** In all the problems considered in the last chapter sets of sufficient statistics exist. This is far from being a general rule. It fails indeed for such a simple form as the Cauchy law

$$P(dx | \alpha, \sigma, H) = \sigma dx / [\pi\{\sigma^2 + (x - \alpha)^2\}].$$

If we have  $n$  observations the likelihood is not capable of being expressed in terms of the unknowns  $\alpha, \sigma$  and any two functions of the observed values of the  $x$ 's. For most of the Pearson laws there are no sufficient statistics. The method of maximum likelihood is applicable to such cases, but is liable to be very laborious, since  $\log L$  must be worked out numerically for at least three trial values of each parameter so that its second derivatives can be found. The result has been, to a very large extent, that where sufficient statistics do not exist for the actual law, it is replaced by one for which they do exist, and information is sacrificed for the sake of ease of manipulation. There is a definite need,

† *J. Inst. Actuaries*, 1947, 1-28.

therefore, for a convenient approximate method that will not lose much of the accuracy given by maximum likelihood but will be reasonably expeditious.

In practice, with almost every method, observations are grouped by ranges of the argument before treatment. Thus effectively the data are not the individual observations but the numbers in assigned groups. Suppose then that the number in a group is  $n_r$ , and the total number  $N$ . According to the law to be found the expectation in the group is  $m_r$ , and

$$\sum m_r = \sum n_r = N. \quad (1)$$

$m_r/N$  is the chance, according to the law, that an observation will fall in the  $r$ th group, and is a calculable function of the parameters in the law. Then the joint chance of  $n_1$  observations in the first group,  $n_2$  in the second, and so on, is

$$L = \frac{N!}{\prod (n_r!)} \prod \left(\frac{m_r}{N}\right)^{n_r} = \frac{N!}{\prod (n_r!)} \prod \left(\frac{n_r}{N}\right)^{n_r} \prod \left(\frac{m_r}{n_r}\right)^{n_r}. \quad (2)$$

The  $m_r$  are the only unknown quantities in this expression, and only the last factor involves their variations. Now put

$$m_r = n_r + a_r N^{1/2}, \quad (3)$$

where  $|a_r N^{1/2}| < n_r$ , and where

$$\sum a_r = 0. \quad (4)$$

Then

$$\begin{aligned} \log L &= \text{constant} + \sum n_r \log \left(1 + \frac{a_r N^{1/2}}{n_r}\right) \\ &= \text{constant} + \sum n_r \left(\frac{a_r N^{1/2}}{n_r} - \frac{a_r^2 N}{2n_r^2}\right) + O(N^{-1/2}) \\ &= \text{constant} - \sum \frac{Na_r^2}{2n_r} \\ &= \text{constant} - \frac{1}{2} \sum \frac{(m_r - n_r)^2}{n_r}, \end{aligned} \quad (5)$$

since the first order terms cancel by (4). Hence, apart from an irrelevant constant, we have

$$\log L = -\frac{1}{2} \chi'^2 = - \sum \frac{(m_r - n_r)^2}{2n_r}. \quad (6)$$

$\chi'^2$  differs from Pearson's  $\chi^2$  only in having  $n_r$  in the denominator instead of  $m_r$ . The difference will be of order  $(m_r - n_r)^3/n_r^2$ , which is of the order of the cubic terms neglected in both approximations. But this form has the advantage that the  $n_r$  are known, while the  $m_r$  are not. We can write the observed frequencies as equations of condition

$$m_r = n_r \pm \sqrt{n_r} \quad (7)$$



and then solve for the parameters in  $m_r$  by the method of least squares, with known weights. Pearson's form is equivalent to this accuracy—it is itself an approximation to  $-2 \log L$ , apart from a constant—but would require successive approximation in actual use on account of the apparent need to revise the  $m_r$  at each approximation. It does not appear that minimum  $\chi^2$  has actually been much used in practice, possibly for this reason. There are some references in the literature to the fitting of frequencies by 'least squares', but the weights to be used are not stated and it is not clear that minimum  $\chi^2$  is meant. The errors due to treating all values of  $n_r$  as having the same accuracy would be serious. The present form was given by Dr. J. Neyman† and rediscovered by myself,‡ Neyman's paper having apparently attracted little attention in this country. The great difficulty in calculating  $\log L$  completely is that it usually requires the retention of a large number of figures; in actual cases  $\log_{10} L$  may be  $-200$  to  $-600$ , and to find the *standard errors to two figures requires that the second decimal should be correct*. But in this method most of  $\log L$  is absorbed into the irrelevant additive constant, and we have only to calculate the changes of the  $m_r$ , given  $N$ , for a set of given small changes of the parameters.

The method fails if any of the  $n_r$  are zero, and is questionable if any of them are 1. For unit groups there appears to be no harm in writing

$$m_r = 1 \pm 1 \quad (8)$$

because if a parameter depends on a single unit group it will be uncertain by its full amount in any case; while if it depends on  $p$  unit groups the equations derived by using (8) for each can be summarized by

$$\sum m_r = p \pm \sqrt{p}, \quad (9)$$

which is right. But special attention is needed for empty groups. Referring to (2) we see that if  $n_r = 0$ ,  $(m_r/N)^{n_r} = 1$  for all values of  $m_r$ . If  $M$  is the sum of the values of  $m_r$  over the empty groups, we can still make the substitution (3), but we shall now have

$$\sum N^{1/2} a_r = -M, \quad (10)$$

$$\log L = \text{constant} - \frac{1}{2} \sum \frac{(m_r - n_r)^2}{n_r} - M, \quad (11)$$

where the summations are now over the occupied groups. Hence if there are empty groups we can take

$$\chi'^2 = \sum \frac{(m_r - n_r)^2}{n_r} + 2M, \quad (12)$$

† *Bull. Inst. Intern. de Statistique*, Warsaw, pp. 44–86 (1929).

‡ *Proc. Camb. Phil. Soc.* 34, 1938, 156–7.

the summation being over the occupied groups, and  $M$  being the total expectation according to the law in the empty groups. The term  $-M$  in  $\log L$  corresponds to the probability  $e^{-r}$  for a zero result according to the Poisson law. This form does not lend itself to immediate solution by least squares. In practice, with laws that give a straggling tail of scattered observations with some empty groups, it is enough to group them so that there are no empty groups,  $m_r$ , for a terminal group being calculated for a range extending to infinity. Then (7) can always be used.†

**4.2. Least square equations : successive approximation.** It often happens that a large number of the coefficients in the normal equations are small or zero. In the extreme case, where all coefficients not in the leading diagonal vanish, the equations are said to be orthogonal. In the other extreme, where the determinant of the coefficients vanishes, the solution is indeterminate, at least one unknown being capable of being assigned arbitrarily. In all intermediate cases the determinant is less than the product of the diagonal elements; if it is much less, the solution may be called badly determined. The solution can, in theory, always be completed on the lines of 3.5, but it often happens that there are, effectively, so many unknowns that it is desirable to do the work piecemeal. Two methods of successive approximation are often suitable.

Consider the form

$$2W = b_{11}x_1^2 + 2b_{12}x_1x_2 + b_{22}x_2^2 + \dots - 2d_1x_1 - 2d_2x_2 - \dots + e, \quad (1)$$

and the normal equations

$$b_{11}x_1 + b_{12}x_2 + \dots = d_1, \quad (2)$$

$$b_{12}x_1 + b_{22}x_2 + \dots = d_2, \quad (3)$$

. . . . .

We can proceed by the following method, due to von Seidel. In (2) neglect all terms in  $x_2, \dots$  and take, therefore,  $x_1 = d_1/b_{11}$ . Now if all the  $x$ 's are 0,  $2W = e$ . If we take  $x_1 = d_1/b_{11}$  and all the others 0,

$$2W = \frac{d_1^2}{b_{11}} - \frac{2d_1^2}{b_{11}} + e, \quad (4)$$

so that this substitution always reduces  $W$ . Now make this substitution in (3) and neglect  $x_3, x_4, \dots$ . Then we have the approximation

$$b_{22}x_2 = d_2 - b_{12}d_1/b_{11}, \quad (5)$$

and  $W$  is reduced by a further amount

$$\frac{1}{b_{22}} \left( d_2 - \frac{b_{12}d_1}{b_{11}} \right)^2. \quad (6)$$

† For numerical illustrations see *Ann. Eugen.* 11, 1941, 108-14.

So we may proceed, substituting in each equation the approximations already found. On reaching the end we begin again at the first equation, using the first approximations for  $x_2$  to  $x_n$ . Since  $W$  is diminished each time the process must converge, and often does so very rapidly. An analogous method has been given recently by R. V. Southwell and A. N. Black under the name of the progressive relaxation of constraints, from an analogy with problems of elasticity.<sup>†</sup>

The following method is sometimes quicker but does not necessarily converge. Begin by transferring all terms of the normal equations to the right side, except the diagonal terms, thus:

$$b_{11}x_1 = d_1 - b_{12}x_2 - b_{13}x_3 - \dots, \quad (7)$$

$$b_{22}x_2 = d_2 - b_{12}x_1 - b_{23}x_3 - \dots, \quad (8)$$

The first approximations are  $x_1 = d_1/b_{11}$ ,  $x_2 = d_2/b_{22}$ , and so on. Substitute on the right to obtain a second approximation, and proceed. Failure of the method will be indicated by failure of the approximations to tend to a limit. In both methods it is a saving of trouble to make a preliminary table of all the ratios  $b_{12}/b_{11}$ ,  $b_{12}/b_{22}$ ,... so as to be able to give at once the correction to any unknown due to a change in any other.

Evidently the rate of convergence in both cases will depend on the latter set of ratios. As an example consider a set of equations

$$\left. \begin{aligned} x_1 &= 1 - kx_2 - kx_3, \\ x_2 &= -kx_1 - kx_3, \\ x_3 &= -kx_1 - kx_2. \end{aligned} \right\} \quad (9)$$

The second method gives  $(1, 0, 0)$  as the first approximation,  $(1, -k, -k)$  as the second,  $(1 + 2k^2, -k + k^2, -k + k^2)$  as the third, and so on. The second approximation always decreases  $W$ , the third decreases it if  $-0.39 < k < 0.64$  but otherwise increases it.

Seidel's method, applied to the same set of equations, gives in turn

$$\begin{aligned} x_1 &= 1, & x_2 &= -k, & x_3 &= -k + k^2, \\ x_1 &= 1 + 2k^2 - k^3, & x_2 &= -k + k^2 - k^3 + k^4, & \dots \end{aligned}$$

The correct solution, to order  $k^3$ , is

$$x_1 = 1 + 2k^2 - 2k^3, \quad x_2 = x_3 = -k + k^2 - 3k^3. \quad (10)$$

The chief usefulness of these methods is in the estimation of many

<sup>†</sup> *Proc. Roy. Soc. A*, **164**, 1938, 447-67; *Relaxation Methods in Engineering Science*, 1940; *Relaxation Methods in Theoretical Physics*, 1946.

unknowns when some of them occur in only a small fraction of the equations of condition. The method of Southwell and Black has been applied, for instance, by the Ordnance Survey to problems where the work is laid out in many stages.† Each point gives rise to equations of condition connecting its position with those of the points observed from it and those it is observed from. Any displacement of its adopted position appears in no equation of condition for a point two stages away, or more, and most of the coefficients in the normal equations are therefore zero. Hence the points can be adjusted in turn, beginning with those observed from the base-line. A modification of the second method was used by Bullen and me in the construction of the times of the  $P$  wave in seismology.‡ Here for each earthquake used there were three special parameters, namely, the latitude and longitude of the epicentre and the time of occurrence. The other parameters to be found were a set of corrections to the trial table at such intervals that interpolation would be possible. What was done was to use the trial tables to determine the elements of each earthquake as if the tables were right. The residuals were then classified by distance to give corrections to the tables. The process was then repeated with the corrected tables as a standard. No change was needed after the third approximation. One advantage of these methods is that they are iterative and therefore self-checking; another is that they break up the work into parts and avoid the need to form and solve what would in this case have been normal equations for about 150 unknowns. The difference from the simple statements of the rules given above is that two or three unknowns are adjusted at once instead of only one.

An estimate of uncertainty can be obtained as follows. Remembering that the standard error of  $x_1$  is  $\sigma(B_{11}/D)^{1/2}$  and that  $B_{11}/D$  is the value found for  $x_1$  on putting 1 on the right of the normal equation for  $x_1$  and 0 in all the others, we need only make this substitution, solve by iteration for each parameter in turn, and the standard errors follow at once.

**4.21. Combination of estimates with different estimated uncertainties.** We have seen that when a set of observations is derived from the normal law, but the standard error is estimated from the residuals, its uncertainty makes the posterior probability of the true value follow the  $t$  rule instead of the normal law. The effect is fully taken into account in the standard tables for the  $t$  rule. But it often happens that several series of observations yield independent estimates of the same true value, the standard errors of one observation being

† *The Observatory*, 62, 1939, 43.

‡ *Bur. Centr. Séism., Trav. Sci.*, Fasc. 11, 1935.

different in the different series. Can we still summarize the information in any useful compact form? The exact solution is straightforward; it is

$$P(dx | \theta H) \propto \prod_r \left\{ 1 + \frac{(x - \bar{x}_r)^2}{\nu_r c_r^2} \right\}^{-1/2(\nu_r + 1)} dx, \quad (1)$$

where  $\bar{x}_r$ ,  $\nu_r$ , and  $c_r$  are the mean, number of degrees of freedom, and standard error of the mean of the  $r$ th set. This can be calculated exactly for any set of estimates, but it is unlikely that the calculation would often be undertaken. Clearly it is in general not reducible to a  $t$  rule.

It would be useful if we could reduce (1) approximately to a  $t$  rule. We are mostly concerned with errors not more than a few times the standard error of our estimate. Consequently it is better to try to fit a  $t$  rule for small errors than for large ones. We can proceed by equating first, second, and fourth derivatives of the logarithms at the value of  $x$  that makes the density in (1) a maximum. It is obviously useless to equate third derivatives, because the  $t$  rule is always symmetrical and (1) need not be exactly so. We try therefore to choose  $\bar{x}$ ,  $c$ ,  $\nu$  so that

$$\frac{1}{2}(\nu + 1) \log \left\{ 1 + \frac{(x - \bar{x})^2}{\nu c^2} \right\} - \frac{1}{2} \sum (\nu_r + 1) \log \left\{ 1 + \frac{(x - \bar{x}_r)^2}{\nu_r c_r^2} \right\} \quad (2)$$

has zero first, second, and fourth derivatives at  $x = \bar{x}$ . The conditions are

$$\sum \frac{\nu_r + 1}{\nu_r c_r^2} \frac{\bar{x} - \bar{x}_r}{u_r(\bar{x})} = 0, \quad (3)$$

$$\frac{\nu + 1}{\nu c^2} = \sum \frac{\nu_r + 1}{\nu_r c_r^2} \left\{ \frac{2}{u_r^2(\bar{x})} - \frac{1}{u_r(\bar{x})} \right\}, \quad (4)$$

$$\frac{\nu + 1}{\nu^2 c^4} = \sum \frac{\nu_r + 1}{\nu_r^2 c_r^4} \left\{ \frac{1}{u_r^2(\bar{x})} - \frac{8}{u_r^3(\bar{x})} + \frac{8}{u_r^4(\bar{x})} \right\}, \quad (5)$$

where 
$$u_r(\bar{x}) = 1 + \frac{(\bar{x} - \bar{x}_r)^2}{\nu_r c_r^2}. \quad (6)$$

These can be solved by successive approximation without much difficulty. It may be noticed that for a single  $t$  rule the expectation of  $1/u_r(\bar{x})$  is  $\nu_r/(\nu_r + 1)$  and that of the right side of (4) is  $\sum (\nu_r + 1)/(\nu_r + 3)c_r^2$ . Hence in a first approximation we can weight the  $\bar{x}_r$  in accordance with their unmodified standard errors, but  $c^{-2}$  will be systematically less than  $\sum c_r^{-2}$ . The approximation therefore corrects the underestimate of the second moment made by using the normal law instead of the  $t$  law for the separate series. The solution allows series even with  $\nu_r = 1$  to be taken into account (cf. 3.4 (13)).  $\nu$  can be called the effective number of degrees of freedom.

In some cases (1) may have more than one maximum. Attempts to combine the estimates are then undesirable.

**4.3. The use of expectations.** When a law of chance is such that sufficient statistics do not exist, it is often possible to proceed by considering some function or functions of the observations. Given the parameters in the law, the expectations of these functions may be calculable in terms of the parameters. But the observations themselves yield the actual values of the functions for that set of observations. If the number of functions is also the number of parameters in the law, estimates of the parameters can be got by equating the theoretical and observed values. If the functions chosen are such that their expectations are actually equal to the parameters they are called unbiased statistics by E. S. Pearson and J. Neyman.

There are apparently an infinite number of unbiased statistics associated with any law. For we might choose any function of the observations, work out its expectation in terms of the law, and transform the law so as to introduce that expectation as a parameter in place of one of the original ones. A choice must therefore be made.

If  $\alpha, \beta, \gamma$  are parameters in a law, we can choose functions of a set of  $n$  possible observations  $f(x_1, x_2, \dots, x_n), g(x_1, \dots, x_n), h(x_1, \dots, x_n)$  and work out their expectations  $F, G, H$ , so that these will be functions of  $\alpha, \beta, \gamma$  and will yield three equations for them when applied to an actual set of observations. Actually, however, the observed values will differ somewhat from the expectations corresponding to the correct values of the parameters. The estimates of  $\alpha, \beta, \gamma$  obtained will therefore be  $a, b, c$ , which will differ a little from  $\alpha, \beta, \gamma$ . The choice is then made so that all of  $E(a-\alpha)^2, E(b-\beta)^2, E(c-\gamma)^2$  will be as small as possible.

It should be noticed that an expectation on a law is not necessarily found best by evaluation of the corresponding function of the observations. Suppose, for instance, that we have a set of observations derived from the normal law about 0 and that for some reason we want the expectation of  $x^4$ . This could be estimated as  $\sum x^4/n$  from the actual observations. Its theoretical value is  $3\sigma^4$ . But

$$\begin{aligned} E\left(\frac{\sum x^4}{n} - 3\sigma^4\right)^2 &= E\left(\frac{\sum x^4}{n}\right)^2 - 6\sigma^4 E\left(\frac{\sum x^4}{n}\right) + 9\sigma^8 \\ &= E\left(\sum \frac{x^4}{n}\right)^2 - 9\sigma^8 \\ &= \frac{1}{n^2} E \sum x^8 + \frac{1}{n^2} E(\sum x^4) E(\sum x^4) - 9\sigma^8, \end{aligned}$$

$\Sigma'$  meaning the sum over all values except the one taken to be  $x$  in  $\Sigma$  (all pairs occurring twice in the double summation); and this is

$$= \frac{\mu_8}{n} - \frac{9}{n} \sigma^8 = \frac{96\sigma^8}{n}.$$

On the other hand, we find

$$E\left(\frac{\sum x^2}{n} - \sigma^2\right)^2 = \frac{2\sigma^4}{n},$$

whence 
$$E(3\sigma^4 - 3\overline{x^2})^2 = \frac{72\sigma^8}{n} + O\left(\frac{1}{n^2}\right).$$

Thus three times the square of the mean square deviation is systematically nearer the fourth moment of the law than the mean of the fourth powers of the deviations is. We should be entitled to call  $\sum x^4/n$  an unbiased statistic for the fourth moment of the law; but it is not the statistic that, given the parameters in the law, would be systematically nearest to the true value. In this case  $\sum x^2/n$  is a sufficient statistic, and we have an instance of the rule that we shall get the best estimates of any function of the parameters in the law by using the sufficient statistics, where these exist.

It may be asked why, seeing that the calculations are done on the hypothesis that  $\sigma$  is known, we should be interested in the probable consequences of taking either  $\overline{x^2}$  or  $\overline{x^4}$  to derive an estimate of  $\sigma$ , seeing that both estimates will be in error to some extent. In this case the interest is not great. The practical problem is usually to estimate  $\sigma$  from the observations, taking the observations as known and  $\sigma$  as initially unknown, and the set of observations is unique. Then we know from the principle of inverse probability that the whole information about  $\sigma$  is summed up in  $\overline{x^2}$  and we need consider no other function of the observations; if we have  $\overline{x^2}$  no other function will tell us anything more about  $\sigma$ , if the normal law is true; if we have not  $\overline{x^2}$ , but have some other function of the scatter of the observations, there must be some loss of accuracy in estimating  $\sigma$ , since  $\overline{x^2}$  is uniquely determined by the observations but will not be uniquely determined by this other function. Nevertheless occasions do arise where it is convenient to use, to provide an estimate, some function of the observations that is not a sufficient statistic. If sufficient statistics do not exist, the posterior probability distribution for a parameter may be unobtainable without a numerical integration with regard to the others, and this is often too formidable an undertaking. Then it is worth while to consider some set of statistics that can be conveniently found from the observations.

This involves some sacrifice of information and of accuracy, but we shall still want to know what precision can be claimed for the estimates obtained. This will involve finding the probability distribution for the statistics used, given the parameters in the law; and then the principle of inverse probability will still give the probability distribution of the parameters in the law, given these statistics. By considerations similar to those of 4.0 the effect of moderate variations in the prior probability is unimportant. We shall have lost some accuracy, but we shall still know how much we have kept.

Fisher has introduced the convenient term 'efficiency', defined as follows. Let  $\sigma^2(\alpha)$  be the expectation of the square of the error of an estimate, obtained by the method of maximum likelihood or inverse probability, and let  $\sigma'^2(\alpha)$  be the corresponding expectation found by some other method. Then the efficiency of the second estimate is defined to mean the limit of  $\sigma^2(\alpha)/\sigma'^2(\alpha)$  when the number of observations becomes large. In most cases both numerator and denominator are of order  $1/n$ , and the ratio has a finite limit. For the normal law the efficiency of the mean fourth power is  $\frac{3}{4}$ . It may be said that such losses of efficiency are tolerable; an efficiency of  $\frac{3}{4}$  means that the standard error of the estimate is 1.15 times as large as the most accurate method would give, and it is not often that this loss of accuracy will affect any actual decision. Efficiencies below  $\frac{3}{4}$ , however, may lead to serious loss. If we consider what actually will happen, suppose that  $\alpha$  is the true value of a parameter,  $a$  the estimate obtained by the most efficient methods, and  $a'$  that obtained by a less efficient one. Then

$$E(a-\alpha)^2 = \sigma^2(\alpha), \quad E(a'-\alpha)^2 = \sigma'^2(\alpha).$$

But these quantities can differ only because  $a'$  is not equal to  $a$ ; and if both  $a$  and  $a'$  are unbiased, so that

$$E(a-\alpha) = E(a'-\alpha) = 0,$$

we have

$$E(a'-a)^2 = \sigma'^2(\alpha) - \sigma^2(\alpha).$$

If  $a'$  has an efficiency of 50 per cent., so that  $\sigma'(\alpha) = \sqrt{2}\sigma(\alpha)$ ,  $a'$  will habitually differ from  $a$  by more than the standard error of the latter. This is very liable to be serious. No general rule can be given; we have in particular cases to balance accuracy against the time that would be needed for an accurate calculation, but as a rough guide it may be said that efficiencies over 90 per cent. are practically always acceptable, those between 70 and 90 per cent. usually acceptable, but those under 50 per cent. should be avoided.

The reason for using the expectation of the square of the error as



the criterion is that, given a large number of observations, the probability of a set of statistics given the parameters, and that of the parameters given the statistics, is usually distributed approximately on a normal correlation surface; for one parameter and one statistic this reduces to the normal law. The standard error appearing in this will be the expectation that we have considered.

One important case where these considerations arise is that of observations derived from an unknown law of error. Suppose that the law is

$$P(dx | \sigma H) = f\left(\frac{x}{\sigma}\right) \frac{dx}{\sigma} \quad (1)$$

and that the origin is taken so that  $E(x) = 0$ . Let  $E(x^2) = \mu_2$ . We know that the chance of the mean of  $n$  observations is nearly normally distributed about 0 with standard error  $(\mu_2/n)^{1/2}$ .  $\mu_2$  is a determinate function of  $\sigma$ . But in the inverse problem we have to find  $\mu_2$  from the observations, and this may be attempted as follows. Consider  $E\{\sum (x - \bar{x})^2\}$  taken over  $n$  observations. The probability distributions of all the observations separately, given  $\sigma H$ , are independent, and

$$\sum (x - \bar{x})^2 = \sum x^2 - 2 \sum x \cdot \bar{x} + n \bar{x}^2 = \sum x^2 - n \bar{x}^2, \quad (2)$$

$$E\{\sum (x - \bar{x})^2\} = (n-1)\mu_2. \quad (3)$$

Hence  $\frac{\sum (x - \bar{x})^2}{n-1}$  will be an unbiased estimate of  $\mu_2$ . It will not, however,

be the accurate value of  $\mu_2$ , and we proceed to consider its expectation of error. We have

$$\begin{aligned} E[\sum (x - \bar{x})^2 - (n-1)\mu_2]^2 &= E[\{\sum (x - \bar{x})^2\}^2 - 2(n-1)\mu_2 \sum (x - \bar{x})^2 + (n-1)^2\mu_2^2] \\ &= E[\sum (x - \bar{x})^2]^2 - (n-1)^2\mu_2^2 \\ &= E[(\sum x^2 - n\bar{x}^2)^2] - (n-1)^2\mu_2^2 \\ &= E[(\sum x^2)^2 - 2n\bar{x}^2 \sum x^2 + n^2\bar{x}^4] - (n-1)^2\mu_2^2. \end{aligned} \quad (4)$$

$$\text{Now} \quad E(\sum x^2)^2 = E \sum x^4 + E \sum x_1^2 \sum' x_2^2, \quad (5)$$

$\sum'$  denoting summation over all  $x$ 's except  $x_1$ ; the 2 is taken into account by the fact that each pair will appear twice. Hence

$$E(\sum x^2)^2 = n\mu_4 + n(n-1)\mu_2^2; \quad (6)$$

also

$$\begin{aligned} E(n\bar{x}^2 \sum x^2) &= \frac{1}{n} E \sum x_1^2 (x_1 + \sum' x_2)^2 \\ &= \frac{1}{n} E(\sum x^4 + 2 \sum x_1^3 \sum' x_2 + \sum x_1^2 \sum' x_2^2) \\ &= \mu_4 + 0 + (n-1)\mu_2^2, \end{aligned} \quad (7)$$

$$E(n^2\bar{x}^4) = E\frac{1}{n^2}(\sum x)^4 = \frac{1}{n}\mu_4 + \frac{3}{n^2}\sum x_1^2 \sum' x_2^2 = \frac{\mu_4}{n} + \frac{3(n-1)}{n}\mu_2^2 \quad (8)$$

(6 having been replaced by 3 to allow for the double summation). Hence†

$$E[\sum (x-\bar{x})^2 - (n-1)\mu_2]^2 = \frac{(n-1)^2}{n}\mu_4 - (n-1)\left(1 + \frac{3}{n}\right)\mu_2^2. \quad (9)$$

Thus the accuracy of the estimate of the second moment of the law will depend on the fourth moment, that of the fourth on the eighth, and so on. Apparently, therefore, we arrive at no result unless we have the complete set of moments; but only  $n$  independent ones can be found from the observations, and for laws of Types IV and VII the higher moments of the law do not exist. However, this is not so serious as it seems. We are usually interested primarily in the mean and its uncertainty, the latter being of order  $n^{-1/2}$ . But the uncertainty of  $\mu_2$  is also of order  $n^{-1/2}$  if  $\mu_4$  exists; and therefore will affect the uncertainty of the mean by something of order  $n^{-1}$ . Quite a rough estimate of  $\mu_4$  will therefore be enough. We can get this by considering

$$E\{\sum (x-\bar{x})^4\} = E[\sum x^4 - 4 \sum x^3\bar{x} + 6 \sum x^2\bar{x}^2 - 3n\bar{x}^4]. \quad (10)$$

Here 
$$E(\sum x^3\bar{x}) = \frac{1}{n}E \sum x_1^3(x_1 + \sum' x_2) = \mu_4, \quad (11)$$

and we find

$$E\{\sum (x-\bar{x})^4\} = (n-1)\left\{\left(1 - \frac{3}{n} + \frac{3}{n^2}\right)\mu_4 + \left(\frac{6}{n} - \frac{9}{n^2}\right)\mu_2^2\right\}. \quad (12)$$

Given the law, the errors of  $\bar{x}$  and  $\sum (x-\bar{x})^2$  are not necessarily independent. We have

$$\begin{aligned} E[n\bar{x}\{\sum (x-\bar{x})^2 - (n-1)\mu_2\}] &= E\{(\sum x)(\sum x^2 - n\bar{x}^2)\} \\ &= E(\sum x^3) - \frac{1}{n}E(\sum x)^3 = (n-1)\mu_3, \end{aligned} \quad (13)$$

$$E\{\sum (x-\bar{x})^3\} = (n-1)\left(1 - \frac{2}{n}\right)\mu_3. \quad (14)$$

There will therefore be a correlation between the errors of location and scaling if the law is unsymmetrical. With such a law, if  $\mu_3$  is positive, there will be a strong concentration of chance at small negative values of  $x$  and a widely spread distribution over positive values. Thus a negative error of the mean will tend to be associated with a small scatter of the observations and a positive one with a large scatter.

The higher moments in such a case furnish an example of what

† This can also be derived easily from Fisher, *Proc. Lond. Math. Soc.* **30**, 1930, 206.

Fisher calls ancillary statistics, which are not used to estimate the parameters but to throw additional light on their precision. The number of observations is always an ancillary statistic.  $\bar{x}$  and  $\sum (x-\bar{x})^2/(n-1)$  are unbiased statistics for the parameter of location and its standard error, but they sacrifice some information contained in the observations if the law is not normal. According as  $\mu_4$  is more or less than  $3\mu_2^2$ , the estimate of  $\mu_2$  will be less or more accurate than a similar estimate from the same number of observations given the normal law. In the former case the posterior probability for the location parameter will resemble a  $t$  distribution with less than  $n-1$  degrees of freedom, in the latter one with more. If for reasons of convenience, then, we take as our estimate  $\bar{x} \pm \left\{ \frac{\sum (x-\bar{x})^2}{n(n-1)} \right\}^{1/2}$ , as for the normal law, attention to  $\mu_3$  and  $\mu_4$  will recover some of the information concerning the distribution of the chance of large errors.

The correlation between  $\bar{x}$  and  $\sum (x-\bar{x})^2$  is

$$\rho = \frac{E[n\bar{x}\{\sum (x-\bar{x})^2 - (n-1)\mu_2\}]}{[E(n^2\bar{x}^2)E\{\sum (x-\bar{x})^2 - (n-1)\mu_2\}^2]^{1/2}} = \frac{\mu_3}{\left\{ \mu_2 \left( \mu_4 - \frac{n-3}{n-1} \mu_2^2 \right) \right\}^{1/2}} \quad (15)$$

and if we write

$$E(\bar{x})^2 = \sigma_1^2, \quad \{\sum (x-\bar{x})^2 - (n-1)\mu_2\} = (n-1)\mu'_2, \quad E(\mu'_2)^2 = \sigma_2^2, \quad (16)$$

we shall have

$$P(d\bar{x}d\mu'_2 | \sigma H) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{(1-\rho^2)}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{\bar{x}^2}{\sigma_1^2} - \frac{2\rho\bar{x}\mu'_2}{\sigma_1\sigma_2} + \frac{\mu'^2_2}{\sigma_2^2}\right)\right\} d\bar{x}d\mu'_2 \quad (17)$$

with considerable accuracy, and this may be used in place of the likelihood in assessing the posterior probabilities when the location and scale parameters are to be found from the observations.

If  $\mu_4$  is infinite, as for a Type VII law with index 2, the expression (9) is infinite, and it appears that the estimate of  $\mu_2$  will have an infinite uncertainty. This does not prove, however, that the estimate is useless. It means only that the chance of error in  $\mu_2$  is so far from being normally distributed that it has an infinite second moment. The law for it will resemble the Cauchy distribution (index 1); though this has an infinite second moment it is possible to find on it a deviation with the same chance of being exceeded as for any given deviation on the normal law; it does not represent infinite uncertainty. But what will be true is that the chance of large errors in  $\mu_2$  as estimated will fall off less rapidly than it will for finite  $\mu_4$  as  $n$  increases.

The method of expectations sometimes fails completely. Karl Pearson's procedure in fitting his laws was to find the mean of the observed values, and the mean second, third, and fourth moments about the mean. These would be equated to  $Ex$ ,  $E(x-Ex)^2$ ,  $E(x-Ex)^3$ , and  $E(x-Ex)^4$ . This process gives four equations for the parameters in the law, which can then be solved numerically. These moments are not in general sufficient statistics, since the likelihood cannot be expressed in terms of them except in a few special cases. The resulting inaccuracy may be very great. For the Type VII law

$$P(dx | \alpha, m, \sigma, H) \propto \frac{dx}{\{1 + (x-\alpha)^2/2m\sigma^2\}^m}$$

when  $m \leq \frac{5}{2}$ , the expectation of the fourth moment is infinite. The actual fourth moment of any set of observations is finite, and therefore any set of observations derived from such a law would be interpreted as implying  $m \geq \frac{5}{2}$ . For some actual series of observational errors  $m$  is as small as this or nearly so. Pearson does not appear to have allowed for finite  $n$ ; he identified  $\sum (x-\bar{x})^r$  with  $\sum x^r$ , neglecting the error of  $\bar{x}$ . This is usually trivial in practice. But Pearson's delight in heavy arithmetic often enabled him to give results to six figures when the third was in error for this reason and the second was uncertain with any method of treatment. The method of minimum  $\chi'^2$  should give greater accuracy with little trouble; other approximate methods, approaching the accuracy of the method of maximum likelihood at its best, are available for Types II and VII, and for I and IV as long as the asymmetry is not too great†; for Types III and V with known termini, sufficient statistics exist. If the terminus is known to be at 0, the arithmetic and geometric means are sufficient for Type III, the geometric and harmonic means for Type V. For the rectangular law the extreme observations are sufficient statistics in any case.

The property of the extreme observations for the rectangular law can be somewhat generalized. For suppose that the lower terminus is at  $x = \alpha$ , and that

$$P(x < x_1 | \alpha H) = A(x_1 - \alpha)^r \quad (18)$$

for  $x_1 - \alpha$  small. Then the chance that  $n$  observations will all be greater than  $x_1$  is  $\{1 - A(x_1 - \alpha)^r\}^n$ , the differential of which will be the chance that the extreme observation will lie in a range  $dx_1$ . Taking the prior probability of  $\alpha$  uniform, we shall have

$$\begin{aligned} P(d\alpha | x_1 H) &\propto \{1 - A(x_1 - \alpha)^r\}^{n-1} (x_1 - \alpha)^{r-1} d\alpha \\ &\propto (x_1 - \alpha)^{r-1} \exp\{-(n-1)A(x_1 - \alpha)^r\} d\alpha \end{aligned} \quad (19)$$

† *Phil. Trans. A*, 237, 1938, 231-71.

for large  $n$ . For  $r = 1$ , the rectangular law, this makes the expectation of  $x_1 - \alpha$ , given  $x_1$ , of order  $1/n$ ; for  $r < 1$ , corresponding to U-shaped and J-shaped distributions, the expectation falls off more rapidly than  $1/n$ ; even for  $r = 2$ , it still only falls off like  $n^{-1/2}$ . Thus even for laws that cut the axis at a finite angle the extreme observation may contain an amount of information about the terminus comparable with that in the remainder; for other laws between this and the rectangular law, and for all U-shaped and J-shaped distributions, the extreme observation by itself may be used to provide an estimate of the terminus. This remark, due originally to Fisher, shows the undesirability of grouping the extreme observations in such cases. It may easily happen that the grouping interval is more than the uncertainty derivable from the extreme observation alone, and then grouping may multiply the uncertainty attainable several times.

Sometimes a law would possess sufficient statistics if certain minor complications were absent. It is then often sufficiently accurate to find expectations of the contributions to these statistics made by the minor complications, and subtract them from the values given by the observations. The method of maximum likelihood can then be used. An example of a common type is given in 4.6.

**4.31. Orthogonal parameters.** It is sometimes convenient to choose the parameters in a law so that the product terms in  $\phi_2$  of 4.0 (4) will have small coefficients. If the maximum likelihood estimates of the parameters in a law  $g(x, \alpha_i)$  are  $a_i$ , and if  $\alpha_i - a_i = \alpha'_i$ ,

$$\begin{aligned} \log L &= \sum_r \log g(x_r, \alpha_i) \\ &= \sum_r \log g(x_r, a_i) + \frac{1}{2} \sum_r \frac{\partial^2}{\partial \alpha_i \partial \alpha_k} \log g \cdot \alpha'_i \alpha'_k, \end{aligned} \quad (1)$$

the derivatives being evaluated at  $\alpha_i = a_i$ . Now the expectation of the coefficient of  $\alpha'_i \alpha'_k$  is

$$\frac{1}{2} n \int_{-\infty}^{\infty} g \frac{\partial}{\partial \alpha_i} \frac{1}{g} \frac{\partial g}{\partial \alpha_k} dx = \frac{1}{2} n \int_{-\infty}^{\infty} \left( -\frac{1}{g} \frac{\partial g}{\partial \alpha_i} \frac{\partial g}{\partial \alpha_k} + \frac{\partial^2 g}{\partial \alpha_i \partial \alpha_k} \right) dx. \quad (2)$$

Since  $\int g dx = 1$  for all  $\alpha_i$ , the second part of the integral is zero; hence

$$E \frac{1}{2} \sum_r \frac{\partial^2}{\partial \alpha_i \partial \alpha_k} \log g \cdot \alpha'_i \alpha'_k = -\frac{1}{2} n g_{ik} \alpha'_i \alpha'_k, \quad (3)$$

where  $g_{ik}$  is the same function as in 3.9. There is therefore a direct relation between the expectation of the quadratic terms in  $\log L$  and the invariant forms  $I_2$  and  $J$  used in 3.9.

Now if  $g_{ik} d\alpha_i d\alpha_k$  is regarded as the square of an element of distance in  $m$  dimensions, at any point it will be possible to choose in an infinity of ways a set of  $m$  mutually orthogonal directions. We can then choose orthogonal coordinates  $\beta_j$ , so that if

$$g_{ik} d\alpha_i d\alpha_k = h_{jl} d\beta_j d\beta_l \quad (4)$$

all  $h_{jl}$  vanish except for  $j = l$ . If the law  $g(x, \alpha_i)$  is then expressed in terms of the quantities  $\beta_j$  instead of  $\alpha_i$ , the quadratic terms in  $E(\log L)$  will reduce to a sum of squares, and for an actual set of observations the square terms in  $\log L$  will increase like  $n$ , while the product terms will be of order  $n^{1/2}$ . Thus the equations to determine the  $\beta_j$  will be nearly orthogonal, and practical solution will be much simplified. The product terms can be neglected for large  $n$ , since their neglect only introduces errors of order  $n^{-1}$ .

For instance, take a Type VII law in the form

$$y = \frac{(m-1)!}{(2\pi M)^{1/2} (m-\frac{3}{2})! \sigma} \left\{ 1 + \frac{(x-\lambda)^2}{2M\sigma^2} \right\}^{-m}, \quad (5)$$

where  $M$  is a function of  $m$ . Evidently

$$\int y \frac{\partial^2}{\partial \lambda \partial \sigma^2} \log y \, dx \quad \text{and} \quad \int y \frac{\partial^2}{\partial \lambda \partial m} \log y \, dx$$

are zero. The condition that  $\int y \frac{\partial^2}{\partial m \partial \sigma^2} \log y \, dx$  shall vanish is found to be

$$\frac{1}{M} \frac{dM}{dm} = \frac{m+1}{m(m-\frac{1}{2})} = \frac{3}{m-\frac{1}{2}} - \frac{2}{m}. \quad (6)$$

For  $y$  to tend to the normal form with standard error  $\sigma$  when  $m \rightarrow \infty$   $M/m$  must tend to 1; we must therefore have

$$M = (m-\frac{1}{2})^3/m^2 \quad (\frac{1}{2} < m < \infty), \quad (7)$$

so that 
$$y = \frac{m!}{(2\pi)^{1/2} (m-\frac{1}{2})^{1/2} (m-\frac{1}{2})! \sigma} \left\{ 1 + \frac{m^2(x-\lambda)^2}{2(m-\frac{1}{2})^3 \sigma^2} \right\}^{-m}. \quad (8)$$

With the law in this form we can form the maximum likelihood equations for  $\lambda$ ,  $\sigma$ , and  $m$ , neglecting non-diagonal terms, and approximation is rapid, any error being squared at the next step.

For Type II laws the corresponding form is

$$y = \frac{(m-\frac{1}{2})!}{\{2\pi(m+\frac{1}{2})\}^{1/2} (m-1)! \sigma} \left\{ 1 - \frac{m^2(x-\lambda)^2}{2(m+\frac{1}{2})^3 \sigma^2} \right\}^m \quad (1 < m < \infty). \quad (9)$$

For  $m \leq 1$ ,  $dy/dx$  does not tend to 0 at the termini. It is then best to take the termini explicitly as parameters.

Specimen curves for  $\lambda = 0$ ,  $\sigma = 1$  are given in the diagram, Fig. 2.

The maximum likelihood equations for a Type VII law in this form are

$$\frac{\partial}{\partial \lambda} \log L = \frac{m}{M\sigma^2} \sum \frac{x-\lambda}{1+(x-\lambda)^2/2M\sigma^2} = 0, \quad (10)$$

$$-\frac{\partial}{\partial \sigma} \log L = \frac{n}{\sigma} - \frac{m}{M\sigma^3} \sum \frac{(x-\lambda)^2}{1+(x-\lambda)^2/2M\sigma^2} = 0, \quad (11)$$

$$-\frac{\partial}{\partial \mu} \log L = nm^2 \left\{ \frac{d}{dm} \log m! - \frac{d}{dm} \log(m-\frac{1}{2})! - \frac{1}{2(m-\frac{1}{2})} \right\} - \sum m^2 \log \left\{ 1 + \frac{(x-\lambda)^2}{2M\sigma^2} \right\} + \sum \frac{m^2(m+1)(x-\lambda)^2}{2\sigma^2(m-\frac{1}{2})^4 \{1+(x-\lambda)^2/2M\sigma^2\}} = 0, \quad (12)$$

where  $\mu = 1/m$ .

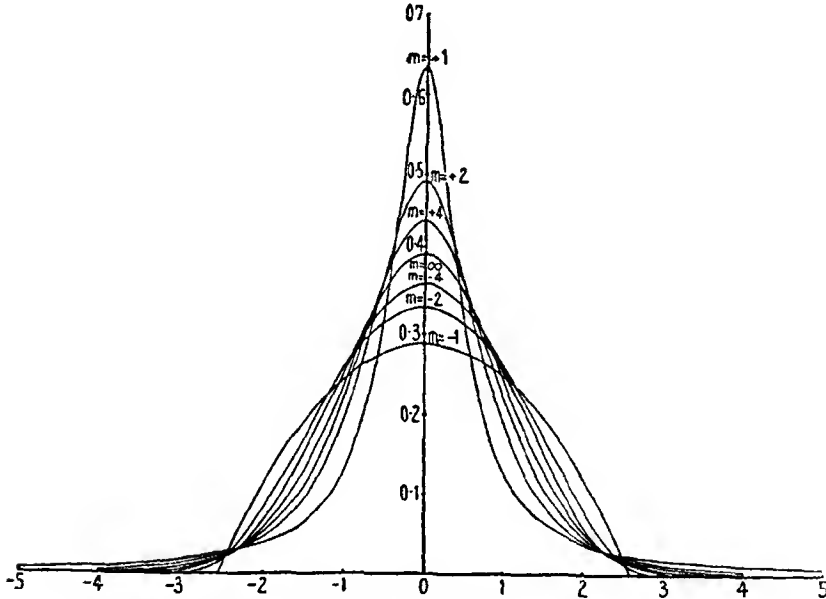


FIG 2. Laws of Types II ( $m$  negative) and VII ( $m$  positive) for  $\sigma = 1$

For Type II they are

$$\frac{\partial}{\partial \lambda} \log L = \frac{m}{M\sigma^2} \sum \frac{x-\lambda}{1-(x-\lambda)^2/2M\sigma^2} = 0, \quad (13)$$

$$-\frac{\partial}{\partial \sigma} \log L = \frac{n}{\sigma} - \frac{m}{M\sigma^3} \sum \frac{(x-\lambda)^2}{1-(x-\lambda)^2/2M\sigma^2} = 0, \quad (14)$$

$$\frac{\partial}{\partial \mu} \log L = nm^2 \left\{ \frac{d}{dm} \log(m-\frac{1}{2})! - \frac{d}{dm} \log(m-1)! - \frac{1}{2(m+\frac{1}{2})} \right\} + \sum m^2 \log \left\{ 1 + \frac{(x-\lambda)^2}{2M\sigma^2} \right\} + \sum \frac{m^2(m-1)(x-\lambda)^2}{2(m+\frac{1}{2})^4 \sigma^2 \{1-(x-\lambda)^2/2M\sigma^2\}} = 0, \quad (15)$$

where  $\mu = -1/m$ . It is convenient to define  $\mu$  as  $+1/m$  for Type VII and as  $-1/m$  for Type II, since this provides for continuous passage through the normal law by increasing  $\mu$  through 0. Actual fitting would be done as follows. First treat  $m$  as infinite and find first approximations to  $\lambda$  and  $\sigma$  as for the normal law. Substitute in (12) or (15), for a number of trial values of  $m$ . Interpolation gives a value of  $\mu$ , and the divided differences give a value of  $\partial^2 \log L / \partial \mu^2$ , which is  $1/s^2(\mu)$ . Return to (10) and (11), or (13) and (14), and derive estimates of  $\lambda$  and  $\sigma$ . If the changes are considerable, solve afresh the equation for  $m$ .

An approximate allowance for the effect of the uncertainty of  $\sigma$  on the posterior probability distribution of  $\lambda$  can be found as follows. For the normal law

$$\left\{ \frac{\partial^2}{\partial \sigma^2} (-\log L) \right\}_{\sigma=s} = \frac{2n}{s^2}.$$

The numerical solution here gives a value for  $s^2(\sigma)$ ; we can define

$$n' = s^2/2s^2(\sigma) = \frac{1}{2}s^2 \left\{ \frac{\partial^2}{\partial \sigma^2} (-\log L) \right\}_{\sigma=s},$$

and, since two parameters besides  $\lambda$  have been estimated, we can take the effective number of degrees of freedom as  $n' - 2$ .

A table of  $d \log m! / dm$  is given by E. Pairman at intervals of 0.02 up to  $m = 20$ .† For  $m \geq 10$  it is given in the British Association Tables.

4.4. If the law of error is unknown and the observations are too few to determine it, we can use the median observation as a statistic for the median of the law. We can then proceed as follows. Let  $\alpha$  be the median of the law; we want to find a range such that the probability, given the observations, that  $\alpha$  lies within it has a definite value. Let  $a$  be a possible value of  $\alpha$  such that  $l$  observations exceed  $a$  and  $n-l$  fall short of it. Then

$$P(l | \alpha, n, H) = {}^nC_l \left(\frac{1}{2}\right)^n = \left(\frac{2}{\pi n}\right)^{1/2} \exp\left\{-\frac{2(l-\frac{1}{2}n)^2}{n}\right\} \quad (1)$$

nearly; and if the prior probability of  $\alpha$  is taken uniform,

$$P(d\alpha | l, n, H) \propto \left(\frac{2}{\pi n}\right)^{1/2} \exp\left\{-\frac{2(l-\frac{1}{2}n)^2}{n}\right\}. \quad (2)$$

Thus the posterior probability density of  $\alpha$  is a maximum at the median, and if we take  $l = \frac{1}{2}n \pm \frac{1}{2}\sqrt{n}$  as limits corresponding to the standard error, the corresponding values of  $\alpha$  will give a valid uncertainty, whatever the law and the scale parameter. The limits will not in general correspond to actual observations but can be filled in by interpolation.

The question of departure from the normal law is commonly

† *Tracts for Computers*, No. 1.



considered in relation to the 'rejection of observations'. Criteria for the latter have been given by Peirce and Chauvenet. They appear, however, to be wrong in principle. If observations are legitimately rejected, the normal law does not hold, and these observations could be used to estimate a departure from it. There is no reason to suppose that the retained observations are themselves derived from the normal law, and, in fact, there is reason to suppose that they are not; and then the mean and standard error found from the observations retained may easily be invalid estimates. Another consideration is that if we make a definite rule that observations within certain arbitrary limits are to be retained at full weight and all beyond them rejected, then the decision about a single outlying observation may easily affect the mean by its apparent standard error, which is highly undesirable. Again it is often advocated that the uncertainty of the true value, as estimated from the mean, should be got from the average residual without regard to sign instead of the mean square residual, on the ground that the former is less affected by a few abnormally large residuals than the latter is. But if the mean of the observations is taken as the estimate of the mean of the law, its uncertainty is correctly estimated from  $\mu_2$ , if the latter exists, and if it does not exist the uncertainty will not be proportional to  $n^{-1/2}$ . For all laws such that  $\mu_2$  exists the mean square residual gives an unbiased estimate of  $\mu_2$ . The ratio of the expectation of the average residual without regard to sign to  $\sqrt{\mu_2}$ , however, depends on the form of the law of error. If the average residual is found and then adapted to give an estimate of  $\sqrt{\mu_2}$  by applying the factor found for the normal law, this factor will be too small for laws of Type VII, which are precisely those where the use of this method is recommended. The cases where this treatment is recommended are just those where it is most likely to lead to an underestimate of uncertainty. If the mean is taken as the estimate, there is no alternative to the mean square residual to provide an estimate of uncertainty when the law is in doubt.

On the other hand, it is only for the normal law that the mean is actually the best estimate, and for other laws we are entitled to consider other estimates that may be more efficient. One interesting case is the law

$$P(dx | m, a, H) = \frac{1}{2} \exp\left(-\frac{|x-m|}{a}\right) \frac{dx}{a}. \quad (3)$$

Here we find easily that the likelihood is a maximum if  $m$  is taken equal to the median observation, and if  $a$  is the average residual without regard to sign. This law is therefore known as the median law. Given

any of the three properties the other two can be deduced. It is only subject to this law that the average residual leads to the best estimate of uncertainty, and then the best estimate of the location parameter is provided by the median observation and not by the mean. The interest of the law is reduced somewhat by the fact that there do not appear to be any cases where it is true. It has the property, however, that it lies higher on the tails and in the centre, and lower on the flanks, than the normal law with the same second moment, and these properties are shared by the laws of Type VII. Fisher shows that for the Cauchy law the standard error of the median of  $n$  observations is  $\pi/2\sqrt{n}$ , while that of the maximum likelihood solution is  $\sqrt{(2/n)}$ . Thus the efficiency of the median as an estimate is  $8/\pi^2 = 0.81$ , which is quite high, in spite of the fact that the expectation of the average residual without regard to sign is infinite. For the normal law it is  $2/\pi = 0.64$ , and it varies little in the intermediate range of Type VII. In the corresponding range the efficiency of the mean varies from 1 to 0. There is, therefore, much to be said for the use of the median as an estimate when the form of the law is unknown; it loses some accuracy in comparison with the best methods, but the increase of the uncertainty is often unimportant, and varies little with the form of the law, and the uncertainty actually obtained is found easily by the rule (2). An extension to the fitting of equations of condition for several unknowns, however, would be rather complicated in practice. The maximum likelihood for the median law comes at a set of values such that, for each unknown, the coefficient of that unknown and the residual have the same sign in half the equations of condition and opposite signs in the other half. To satisfy these relations would apparently involve more arithmetic than the method of least squares. The simplicity of the use of the median for one location parameter does not persist for several parameters, and the practical convenience of the method of least squares is a strong argument for its retention.

**4.41.** The nature of the effect of the law of error on the appropriate treatment is seen by considering a law

$$P(dx | \alpha H) = f(x - \alpha) dx. \quad (1)$$

The maximum likelihood solution is given by

$$\begin{aligned} 0 &= \frac{d}{d\alpha} \log\{f(x_1 - \alpha)f(x_2 - \alpha)\dots f(x_n - \alpha)\} \\ &= \frac{f'(x_1 - \alpha)}{f(x_1 - \alpha)} + \dots + \frac{f'(x_n - \alpha)}{f(x_n - \alpha)}. \end{aligned} \quad (2)$$

If the arithmetic mean is the maximum likelihood solution for all possible observed values, this is equivalent to

$$0 = (x_1 - \alpha) + \dots + (x_n - \alpha), \quad (3)$$

whence 
$$f(x) = Ae^{-h^2x^2}, \quad (4)$$

the result obtained by Gauss. But if we put

$$\frac{f'(x-\alpha)}{(x-\alpha)f(x-\alpha)} = w, \quad (5)$$

(2) is equivalent to 
$$\sum w(x-\alpha) = 0. \quad (6)$$

Hence  $\alpha$  is a weighted mean of the observed values. If  $f'(x)/f(x)$  does not increase as fast as the residual, the appropriate treatment will give reduced weight to the large residuals. If it increases faster, they should receive more weight than the smaller ones. The former consideration applies to a Type VII law, for which, for large  $x-\alpha$ ,

$$f'(x-\alpha)/(x-\alpha)f(x-\alpha)$$

behaves like  $-(x-\alpha)^{-2}$  instead of being constant. The latter applies to the rectangular law, for which  $w$  is zero except at the ends of the range, where it is infinite.

These considerations suggest an appropriate treatment in cases where the distribution is apparently nearly normal in the centre, but falls off less rapidly at the extremes. This kind of distribution is shown especially by seismological observations. If two observers read ordinary records and agree about which phases to read, they will usually agree within 1 or 2 seconds. But the arrival of a new phase is generally superposed on a background which is already disturbed, and the observer has to decide which new onsets are distinct phases and which are merely parts of the background. The bulk of the observers actually usually agree, but there are scattered readings up to 10 or 20 seconds away from the main concentration. The following are specimens. The residuals are in seconds. The first series refer to *P* at good Pacific stations, the second to intermediate ones, the third to *S* at short distances in deep-focus earthquakes.

Residual	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10
Number (1)	0	1	1	1	1	1	4	8	13	14	13	8	10	2	4	1	1	2	2	0	1
Number (2)	0	1	2	0	1	2	2	2	7	8	10	10	4	3	3	2	4	1	2	0	2
Number (3)	?	?	5	4	7	10	16	23	31	51	59	44	39	22	15	8	8	7	8	?	?

The central groups alone may suggest a standard error of about 2 seconds, but the second moment of the whole of the observations might at the best suggest one of 4 or 5 seconds. At the worst it would become

meaningless because there may be no definite gap between two distinct phases, and we have no rule so far for separating them. In such a case we may suppose that the law has the form

$$P(dx | \alpha, h, H) = \frac{(1-m)h}{\sqrt{\pi}} \exp\{-h^2(x-\alpha)^2\} + mg(x-\beta), \quad (7)$$

where  $mg$  is always small and  $g$  varies little within ranges of order  $1/h$ . Within this range we must regard  $g$  as an unknown function. Then

$$\log L = \sum \log \left[ \frac{(1-m)h}{\sqrt{\pi}} \exp\{-h^2(x-\alpha)^2\} + mg(x-\beta) \right], \quad (8)$$

$$\frac{1}{L} \frac{\partial L}{\partial \alpha} = \sum \frac{\{2(1-m)h^3/\sqrt{\pi}\}(x-\alpha)\exp\{-h^2(x-\alpha)^2\}}{\{(1-m)h/\sqrt{\pi}\}\exp\{-h^2(x-\alpha)^2\} + mg(x-\beta)}, \quad (9)$$

$$\frac{1}{L} \frac{\partial L}{\partial h} = \sum \frac{\{(1-m)/\sqrt{\pi}\}\{1-2h^2(x-\alpha)^2\}\exp\{-h^2(x-\alpha)^2\}}{\{(1-m)h/\sqrt{\pi}\}\exp\{-h^2(x-\alpha)^2\} + mg(x-\beta)}, \quad (10)$$

or, if we write

$$w^{-1} = 1 + \frac{m}{1-m} \frac{\sqrt{\pi}}{h} g(x-\beta) \exp\{h^2(x-\alpha)^2\}, \quad (11)$$

$$\frac{1}{L} \frac{\partial L}{\partial \alpha} = \sum 2h^2 w(x-\alpha), \quad (12)$$

$$\frac{h}{L} \frac{\partial L}{\partial h} = \sum w\{1-2h^2(x-\alpha)^2\}. \quad (13)$$

Thus, with the appropriate weights, the equations for  $\alpha$  and  $h$  reduce to the usual ones. To find these weights requires an estimation of  $g$ , which need only be rough. We note that the density at large residuals is  $mg$ , and at small ones  $(1-m)h/\sqrt{\pi} + mg$ ; thus the coefficient of the exponential in (11) is the ratio of the density at large values to the excess at the mode, which is estimated immediately from the frequencies. If we denote this coefficient by  $\mu$ , we have

$$w^{-1} = 1 + \mu \exp\{h^2(x-\alpha)^2\}, \quad (14)$$

and apart from  $\mu$ ,  $g$  is irrelevant to  $\alpha$  and  $h$ . Also, in  $\partial^2 \log L / \partial \alpha^2$ , the term in  $\partial w / \partial \alpha$  is small on account of a factor  $\mu(x-\alpha)^2$  when  $x-\alpha$  is small, and of a factor  $\mu^{-1} \exp\{-h^2(x-\alpha)^2\}$  when  $x-\alpha$  is large; in any case we can neglect it and take

$$\alpha = \frac{\sum wx}{\sum w} \pm \frac{\sigma}{\sqrt{(\sum w)}}, \quad (15)$$

where  $\sigma$  is given by  $\sigma^2 \sum w = \sum w(x-\alpha)^2$ . (16)

The method has been applied extensively in seismology with satisfactory results. A change in  $h$  or  $\alpha$  necessitates a change of the weights,

and it is usually necessary to proceed by successive approximation, but more often than not the second approximation almost repeats the first. As a starting-point we can find  $h$  roughly from the distributions of the frequencies near the centre, compute from it the expected frequencies according to the normal law, and use the excess on the flanks to estimate  $\mu$ . Alternatively, if there is a range of approximately constant frequencies on each side, we can subtract their mean from *all* frequencies, including those in the central group, replace negative values by 0, and compute  $\sigma$  from the remainders. This has been called the method of *uniform reduction*. The chief use has been in finding corrections to trial tables. The residuals for all ranges together give a good determination of the weights, which are then applied to the separate ranges to give the required corrections. With this method the weight is a continuous function of the residual, and the difficulty about a hard and fast limit for rejection does not arise.

4.42. In the usual statement of the problem of least squares the whole of the uncertainty is supposed concentrated in one of the variables observed, the others being taken as not subject to error. This is a common state of affairs, but not a universal one. It may happen that we have a set of pairs  $(x, y)$ , which may be taken as estimates of two variables  $(\xi, \eta)$  on different occasions, with a linear relation between them, and that the uncertainties of each determination of  $x$  and  $y$  are known and independent. The problem is to find the relation between  $\xi$  and  $\eta$ . Write

$$\eta = \alpha\xi + \beta. \quad (1)$$

Then a typical observation  $(x_r \pm s_r, y_r \pm t_r)$  must be read as

$$P(dx_r dy_r d\xi_r | \alpha, \beta, H) = \frac{1}{2\pi s_r t_r} \exp \left\{ -\frac{(x_r - \xi_r)^2}{2s_r^2} - \frac{(y_r - \alpha\xi_r - \beta)^2}{2t_r^2} \right\} dx_r dy_r d\xi_r \quad (2)$$

$$\text{and} \quad \log L = \text{constant} - \sum \left\{ \frac{(x_r - \xi_r)^2}{2s_r^2} + \frac{(y_r - \alpha\xi_r - \beta)^2}{2t_r^2} \right\}, \quad (3)$$

the unknowns being the various  $\xi_r$ ,  $\alpha$ , and  $\beta$ . Integrating with regard to all the  $\xi_r$  we get, with a uniform prior probability for  $\alpha$  and  $\beta$ ,

$$P(d\alpha d\beta | \theta H) \propto \prod (t_r^2 + \alpha^2 s_r^2)^{-1/2} \exp \left\{ -\sum \frac{(y_r - \alpha x_r - \beta)^2}{2(t_r^2 + \alpha^2 s_r^2)} \right\} d\alpha d\beta. \quad (4)$$

Hence we can write

$$\alpha x_r + \beta = y_r \pm (t_r^2 + \alpha^2 s_r^2)^{1/2}, \quad (5)$$

as a set of equations of condition to determine  $\alpha$  and  $\beta$ . Since the standard error involves  $\alpha$  the solution must be by successive approximation, but if the variation of  $x_r$  and  $y_r$  is much more than that of  $s_r$  and

$t_r$ , a first approximation using equal weights will give a good estimate of  $\alpha$  and the second approximation will need little change. The result is equivalent to using  $x_r$  as the correct value of  $\xi_r$ , but using (1) and  $s_r$ , with an approximate  $\alpha$ , to estimate the uncertainty of  $\eta$  at  $\xi_r = x_r$ .

**4.43. Grouping.** Suppose that we have observations  $x_r$  of a quantity, for  $n$  different values of an argument  $t$ , and that we regard these as representing a linear function of  $t$ , say  $\alpha + \beta t$ ; the standard error of each observation is  $\sigma$ . Then a typical equation of condition will be

$$x_r = \alpha + \beta t_r \quad (1)$$

and the normal equations for  $\alpha$  and  $\beta$  will be

$$n\alpha + \beta \sum t_r = \sum x_r, \quad (2)$$

$$\alpha \sum t_r + \beta \sum t_r^2 = \sum t_r x_r, \quad (3)$$

whence the standard error of  $\beta$  is  $\left\{ \frac{n}{n \sum t_r^2 - (\sum t_r)^2} \right\}^{1/2} \sigma$ . If  $\bar{t}$  is the mean of the  $t_r$ , the standard error of  $\alpha + \beta \bar{t}$  is  $\sigma/\sqrt{n}$ , and these uncertainties are independent. This is the most accurate procedure.

On the other hand, we may proceed by taking the means of ranges of observations near the beginning and the end; the difference will then yield a determination of  $\beta$ . If there are  $m$  in each of these ranges and the means are  $(\bar{t}_1, \bar{x}_1)$ ,  $(\bar{t}_2, \bar{x}_2)$ , we have

$$\bar{x}_1 = \alpha + \beta \bar{t}_1 \pm \sigma/\sqrt{m}, \quad \bar{x}_2 = \alpha + \beta \bar{t}_2 \pm \sigma/\sqrt{m}, \quad (4)$$

whence 
$$\beta = \frac{\bar{x}_2 - \bar{x}_1}{\bar{t}_2 - \bar{t}_1} \pm \left( \frac{2}{m} \right)^{1/2} \frac{\sigma}{\bar{t}_2 - \bar{t}_1}. \quad (5)$$

Let us compare the uncertainties on the hypothesis that the observations are uniformly spaced from  $t = -1$  to  $+1$ . Then  $\sum t_r^2$  will be nearly  $\frac{1}{3}n$ , and the least squares solution has standard error  $\sigma\sqrt{(3/n)}$ . Also  $\bar{t}_2 - \bar{t}_1 = 2(1 - m/n)$  and the solution by grouping has standard error  $\sigma/(2m)^{1/2}(1 - m/n)$ . The latter is a minimum if  $m = \frac{1}{2}n$ , and then is equal to  $\sigma(27/8n)^{1/2}$ . The efficiency of the solution by grouping, as far as  $\beta$  is concerned, is therefore  $\frac{8}{27}$ , which for most purposes would be quite satisfactory.† The expectation of the square of the difference between the two estimates would correspond to a standard error  $\frac{1}{3}$  of that of the better estimate. If we took  $m = \frac{1}{2}n$ , we should get a standard error of  $2\sigma/n^{1/2}$ , and the efficiency would be  $\frac{2}{3}$ .

The best estimate of  $\alpha + \beta \bar{t}$  is the mean observation, and it is of no importance whether we average the observations all together or average the means of the three ranges. Hence we shall sacrifice

† The result is due to Sir Arthur Eddington, but he did not publish it.

hardly any accuracy if we divide the observations into ranges each containing a third of the observations, determine  $\beta$  by comparison of the first and third, and  $\alpha + \beta t$  from the mean of all three with equal weight.

Again, suppose that  $t$  ranges from 0 to  $2\pi$ , and that we want to determine  $\alpha + \beta \cos t$  from the observations of  $x$ . The normal equations are

$$n\alpha + \beta \sum \cos t_r = \sum x_r, \quad (6)$$

$$\alpha \sum \cos t_r + \beta \sum \cos^2 t_r = \sum x_r \cos t_r. \quad (7)$$

If the arguments are equally spaced we shall have  $\sigma^2(\alpha) = \sigma^2/n$ ,  $\sigma^2(\beta) = 2\sigma^2/n$ .

But we may compare means by ranges about 0 and  $\pi$ . The sum of the observations between 0 and  $p\pi$  and between  $(2-p)\pi$  and  $2\pi$  will give, nearly,

$$np\alpha + \frac{n\beta}{2\pi} \int_{-p\pi}^{p\pi} \cos t \, dt = n\bar{x}_1 \pm \sigma\sqrt{(np)} \quad (8)$$

and the corresponding equation for the opposite range follows. Hence  $\beta$  can be estimated from

$$2\frac{\beta}{\pi} \sin p\pi = \bar{x}_1 - \bar{x}_2 \pm \sigma\sqrt{(2p/n)} \quad (9)$$

and will be found most accurately if  $p^{1/2} \operatorname{cosec} p\pi$  is a minimum. This leads to  $p\pi = 66^\circ 47'$ . The convenient value  $p\pi = 60^\circ$  gives

$$\sigma^2(\beta) = \frac{2\pi^2}{9} \sigma^2 \quad (10)$$

and the efficiency is  $9/\pi^2 = 0.91$ . If we take  $p = \frac{1}{2}$ , thus comparing whole semicircles, we get an efficiency of  $8/\pi^2 = 0.81$ . The use of opposite ranges of  $120^\circ$ , while giving high efficiency, also has the merit that any Fourier term whose argument is a multiple of two or three times that of the term sought will contribute nothing to the estimate. If we used ranges of  $180^\circ$ , a term in  $3t$  would contribute to the estimate of  $\beta$ , but this term contributes nothing to the mean in a  $120^\circ$  range.

Thus drastic grouping, if done in the best way, loses little in the accuracy of the estimates. The corresponding analysis for frequencies instead of measures leads to the same results.† There may, however, be serious loss when the chance considered falls off rapidly towards the tails. I found this in discussing errors of observation; the sacrifice of

† *Proc. Roy. Soc. A*, 164, 1938, 311–14.

the information about the distribution of the errors in ranges where the expectations according to the normal law were small led to the standard errors being increased several times.

The method is particularly useful in carrying out harmonic analysis. When the data are measures, if we use opposite ranges of  $120^\circ$ , the coefficient of a sine or cosine is given by

$$\begin{aligned}\beta &= \frac{\pi}{\sqrt{3}}(\bar{x}_1 - \bar{x}_2) \pm \frac{\pi\sigma\sqrt{2}}{3\sqrt{n}} \\ &= 1.814(\bar{x}_1 - \bar{x}_2) \pm 1.481\sigma/\sqrt{n}.\end{aligned}\quad (11)$$

Where the problem is to estimate a Fourier term in a chance, if  $n_1$  and  $n_2$  are the numbers of observations in opposite ranges of  $120^\circ$ , we get

$$\beta = 1.814 \frac{n_1 - n_2}{n} \pm \frac{1.481}{\sqrt{n}}. \quad (12)$$

The similarity of the coefficients corresponds to the result in the minimum  $\chi^2$  approximation that we can enter an observed number in an equation of condition as  $n_r \pm \sqrt{n_r}$ .

**4.44. Effects of grouping: Sheppard's corrections.** In some cases it is desirable to make allowance for less drastic grouping than in 4.43. Suppose, as in 3.41, that the true value is  $x$  and the standard error  $\sigma$ , and that we take a convenient arbitrary point of reference  $x_0$ . Then all observations between  $x_0 + (r \pm \frac{1}{2})h$  will be entered as  $x_0 + rh$ , and our data are the numbers of observations so centred. As before, we can take

$$P(dxd\sigma | H) \propto dxd\sigma/\sigma, \quad (1)$$

but the chance of an observation being given as  $x_0 + rh$  is now

$$P(r | x, \sigma, H) = \frac{1}{\sqrt{(2\pi)}\sigma} \int_{x_0 + (r - \frac{1}{2})h}^{x_0 + (r + \frac{1}{2})h} \exp\left\{-\frac{1}{2} \frac{(\xi - x)^2}{\sigma^2}\right\} d\xi. \quad (2)$$

Two cases arise according as  $h$  is large or small compared with  $\sigma$ . In the former case the chance is negligible except for the range that includes  $x$ . Hence if we find nearly the whole of the observations in a single range we shall infer that  $\sigma$  is small compared with  $h$ . The likelihood is nearly constant for values of  $x$  in this range, and we shall be left with a nearly uniform distribution of the posterior probability of  $x$  within the interval that includes the observations, no matter how many observations we have. This is an unsatisfactory result; the remedy is to use a smaller interval of grouping.



If  $h$  is small with regard to  $\sigma$ , and if we put

$$\xi - x_0 - r\hbar = \eta, \quad (3)$$

$$\begin{aligned} & \int_{-\frac{i\hbar}{2}}^{\frac{i\hbar}{2}} \exp \left[ -\frac{1}{2\sigma^2} \{ (x_0 + r\hbar - x)^2 + 2\eta(x_0 + r\hbar - x) + \eta^2 \} \right] d\eta \\ &= \exp \left[ -\frac{1}{2\sigma^2} (x_0 + r\hbar - x)^2 \right] \int_{-\frac{i\hbar}{2}}^{\frac{i\hbar}{2}} \left[ 1 - \frac{\eta}{\sigma^2} (x_0 + r\hbar - x) + \right. \\ & \quad \left. + \frac{\eta^2}{2\sigma^4} (x_0 + r\hbar - x)^2 - \frac{\eta^2}{2\sigma^2} \right] d\eta \\ &= \exp \left[ -\frac{1}{2\sigma^2} (x_0 + r\hbar - x)^2 \right] h \left[ 1 + \frac{h^2}{24\sigma^4} \{ (x_0 + r\hbar - x)^2 - \sigma^2 \} \right] \quad (4) \end{aligned}$$

to order  $h^3$ ; and we shall have for the joint probability of the observations given  $x$  and  $\sigma$ ,

$$\begin{aligned} P(\theta | x, \sigma, H) &\propto \sigma^{-n} \exp \left[ -\frac{1}{2\sigma^2} \sum (x_0 + r\hbar - x)^2 + \frac{h^2}{24\sigma^4} \{ \sum (x_0 + r\hbar - x)^2 - n\sigma^2 \} \right] \\ &\propto \sigma^{-n} \exp \left[ -\frac{n}{2\sigma^2} \{ (\bar{x} - x)^2 + s^2 \} + \frac{nh^2}{24\sigma^4} \{ (\bar{x} - x)^2 + s^2 - \sigma^2 \} \right], \quad (5) \end{aligned}$$

where  $\bar{x}$  and  $s^2$  are a mean and a mean square residual found from the recorded values. To this accuracy they are still sufficient statistics. Hence

$$\begin{aligned} P(dxd\sigma | \theta H) &\propto \sigma^{-n-1} \exp \left[ -\frac{n}{2\sigma^2} (\bar{x} - x)^2 \left( 1 - \frac{h^2}{12\sigma^2} \right) - \frac{ns^2}{2\sigma^2} \left( 1 - \frac{h^2}{12\sigma^2} \right) - \frac{nh^2}{24\sigma^2} \right] dx d\sigma. \quad (6) \end{aligned}$$

Differentiating (5) or (6) we see that the maximum for  $x$  is at  $\bar{x}$ , and that for  $\sigma$  is at

$$\sigma^2 = s^2 - \frac{1}{12}h^2 + O(n^{-1}). \quad (7)$$

The coefficient of  $(x - \bar{x})^2$  in (6) is therefore, to this order,

$$\frac{n}{2(s^2 - \frac{1}{12}h^2)} \left( 1 - \frac{h^2}{12s^2} \right) = \frac{n}{2s^2}. \quad (8)$$

Without the allowance for finite  $h$  the corresponding values would be  $s^2$  and  $n/2s^2$ . Hence (1) the uncertainty of  $x$  can be taken from the mean square residual as it stands, and needs no correction; (2) to estimate  $\sigma^2$  we should reduce  $s^2$  by  $h^2/12$ .

The latter correction is due to W. F. Sheppard.† He proceeded by considering the expectation of the contribution to  $s^2$ , given  $\sigma$ , due to

† *Proc. Lond. Math. Soc.* 29, 1898, 368.

the finite  $h$ , and obtained the correction in this sense for any law of error. He showed also that the contribution to the third moment is zero, and to the fourth  $\frac{1}{2}h^2s^2 - \frac{1}{10}h^4$ , which should therefore be subtracted from the mean fourth moment of the observations before finding that of the law. It is in this form that the corrections have been most used. But the above argument brings out the point, also made by Fisher, that the uncertainty of the true value, given the observations, is determined by the uncorrected second moment and not by the corrected one. It is only when, as in computing a correlation from grouped data, we are directly interested in  $\sigma^2$ , that there is any point in applying the correction. There will be a slight departure from the rule of 3.41 in the posterior probability distribution of  $x$ , but this is negligible.

4.45. There is a similar complication when the standard error consists of two parts, one of which may be supposed known and equal to  $\sigma'$ , while the other is to be found. There are two plausible assessments of the prior probability. We may take  $\sigma$  to be the complete standard error, but restricted now to be greater than  $\sigma'$ ; then the rule would be

$$P(d\sigma | H) \propto d\sigma/\sigma, \quad (1)$$

for  $\sigma > \sigma'$ . On the other hand, we might take this rule to apply to only the unknown portion  $(\sigma^2 - \sigma'^2)^{1/2}$ ; then

$$P(d\sigma | H) \propto d \log(\sigma^2 - \sigma'^2)^{1/2} \propto \frac{\sigma d\sigma}{\sigma^2 - \sigma'^2}. \quad (2)$$

But the latter leads to an absurd result. For the likelihood is still proportional to

$$\sigma^{-n} \exp \left[ -\frac{n}{2\sigma^2} \{ (x - \bar{x})^2 + s^2 \} \right] \quad (3)$$

and (2) will lead to a pole in the posterior probability at  $\sigma = \sigma'$ . Thus the inference using this assessment of the prior probability would be that  $\sigma = \sigma'$ , even though the maximum likelihood will be at a larger value of  $\sigma$ ; (1) on the other hand leads to the usual rule except for a negligible effect of truncation.

The situation seems to be that in a case where there is a known contribution to the standard error it is not legitimate to treat the rest of the standard error as unknown, because the known part is relevant to the unknown part. The above allowance for grouping is a case in point, since we see that it is only when  $h$  is small compared with  $\sigma$  that  $n$  observations are better than one; if the interval was found too large it would in practice be taken smaller in order that this condition should be satisfied. The case that attracted my attention to the problem was

that of observations of gravity, where repetition of observations at the same place shows that the accuracy of observation is of the order of 3 milligals (1 milligal =  $0.001 \text{ cm./sec.}^2$ ), but there are differences between neighbouring places of the order of 20 to 50 milligals. In combining the data to obtain a representative formula the latter must be treated as random variation, to which the inaccuracy of observation contributes only a small known part. The use of (2) would then say that we shall never dispose of the possibility that the whole of the variation is due to the observational error; whereas it is already disposed of by the comparison of observations in different places with the differences between observations repeated at the same place. This is a case of intraclass correlation (see later, 5.6); we must break up the whole variation into a part between stations and a part between observations at the same station, and when the existence of the former is established the standard error is found from the scatter of the station means, the differences between observations at the same station having little more to say. Thus the proper procedure is to use (1) or else to treat the standard error as a whole as unknown, it does not matter which.

**4.5. Smoothing of observed data.** It often happens that we have a series of observed data for different values of the argument and with known standard errors, and that we wish to remove the errors as far as possible before interpolation. In many cases we already know the form of the function to be found, and we have only to determine the most probable values of the parameters in this function. The best method is then the method of least squares. But there are cases where no definite form of the function is suggested. Even in these the presence of errors in the data is expected. The tendency of random error is always to increase the irregularities, and part of any irregularity can therefore be attributed to random error, and we are entitled to try to reduce it. Such a process is called smoothing. Now it often happens in such cases that most of the third, or even the second or first differences, at the actual tabular intervals, are no larger than the known uncertainty of the individual values will explain, but that the values at wider intervals show these differences to be systematic. Thus if we have values at unit intervals of the argument over a range of 40, and we take differences at intervals 10, any systematic second difference will be 100 times as large as for unit intervals, the random error remaining the same. The situation will be, then, that the values at unit intervals give no useful determination of the second derivative of the function, but

this information can be provided by using wider intervals. On the other hand we want our solution to be as accurate as possible, and isolated values will not achieve this; thus the observed values from argument 15 to 25 will all have something to say about the true value at 20, and we need to arrange our work so as to determine this as closely as we can.

In such a case we may find that the values over a range of 10 are enough to determine a linear function by least squares, but that the coefficient of a square term is comparable with its standard error. If we reject the information about the curvature provided by a range of 10, we lose little; and in any case comparison with adjacent ranges will give a much better determination. This suggests that in a range of 10 we may simply fit a linear function. But if we do this there will be discontinuities wherever the ranges abut, and we do not want to introduce new spurious discontinuities. We notice, however, that a linear function is uniquely determined by two values. If then we use the linear solution to find values for two points in each range we can interpolate through all ranges and retain all the information about the curvature that can be got by comparison of widely separated values; while the result for these two values will be considerably more accurate than for the original ones. Such values may be called *summary values*.

Now the two values of the independent variable may be chosen arbitrarily, in an infinite number of ways consistent with the same linear equation. The question is, which of these is the best? We have two considerations to guide us. The computed values will still have errors, of two types: (1) Even if the function sought was genuinely linear, any pair of values found from the observed ones would have errors. If we take the values of the argument too close together, these errors will tend to be equal; if they are too far apart they will tend to have opposite signs on account of the error of the estimated gradient. There will be a set of pairs of values such that the errors are independent. But any interpolated value is a linear function of the basic ones. If we choose one of these pairs, the uncertainty of any interpolated value can be got by the usual rule for compounding uncertainties, provided that these are independent. If they are not, allowance must be made for the correlation, and this makes the estimation of uncertainty much more difficult. (2) We are neglecting the curvature in any one range, not asserting it to be zero. At some points in the range the difference between the linear solution and the quadratic solution, both by least squares, will be positive, at others negative. If we choose summary values at places where the two solutions agree, they are independent

of the curvature and therefore of its uncertainty; and this will not hold of any others. Neglect of the curvature will therefore do least harm if we use these values. We have therefore, apparently, three conditions to be satisfied by the values chosen for the argument: they must be such that the uncertainties of the estimated values of the function at them are independent, and such that neither is affected by the curvature. There are only two quantities to satisfy these conditions, but it turns out that they can always be found.

Let  $x$  be the independent variable,  $y$  the dependent one. Suppose that the summary values are to be at  $x_1$  and  $x_2$ , where  $y$  takes the values  $y_1$  and  $y_2$ . Then the general quadratic expression that takes these values is

$$y = \frac{y_1(x-x_2)-y_2(x-x_1)}{x_1-x_2} + A(x-x_1)(x-x_2), \quad (1)$$

in which  $y_1$ ,  $y_2$ , and  $A$  can be found by least squares. The weight of the equation of condition for a particular  $x$  being  $w$ , the normal equation for  $y_1$  is

$$\begin{aligned} \frac{\sum w(x-x_2)^2 y_1 - \sum w(x-x_1)(x-x_2) y_2}{(x_1-x_2)^2} + \frac{\sum w(x-x_1)(x-x_2)^2 A}{x_1-x_2} \\ = \frac{\sum w(x-x_2) y}{x_1-x_2}. \end{aligned} \quad (2)$$

The conditions that the uncertainties of  $y_1$ ,  $y_2$ , and  $A$  shall be independent are therefore

$$\sum w(x-x_1)(x-x_2) = 0, \quad (3)$$

$$\sum w(x-x_1)(x-x_2)^2 = 0, \quad (4)$$

$$\sum w(x-x_1)^2(x-x_2) = 0. \quad (5)$$

But if we subtract (5) from (4) and cancel a factor  $x_1-x_2$  from all terms we obtain (3). Hence we have only two independent equations to determine  $x_1$  and  $x_2$  and the problem has a solution.

Put now

$$\sum w = n, \quad \sum wx = n\bar{x}, \quad x - \bar{x} = \xi, \quad \sum w\xi^2 = n\mu_2, \quad \sum w\xi^3 = n\mu_3. \quad (6)$$

Then (3) becomes

$$0 = \sum w(\xi - \xi_1)(\xi - \xi_2) = n(\mu_2 + \xi_1 \xi_2) \quad (7)$$

since  $\sum w\xi = 0$ . Also either of (4) or (5) with this gives

$$\mu_3 - \mu_2(\xi_1 + \xi_2) = 0. \quad (8)$$

Hence  $\xi_1$  and  $\xi_2$  are the roots of

$$t^2 - \frac{\mu_3}{\mu_2}t - \mu_2 = 0, \quad (9)$$

and this is the solution required.

The sum of the weights of  $y_1$  and  $y_2$  is easily seen to be  $n$ . For

$$\begin{aligned} \sum w(x-x_1)^2 + \sum w(x-x_2)^2 &= \sum w(\xi-\xi_1)^2 + \sum w(\xi-\xi_2)^2 \\ &= 2n\mu_2 - 2(\xi_1+\xi_2) \sum w\xi + n(\xi_1^2+\xi_2^2) \\ &= 2n\mu_2 + n\{(\xi_1+\xi_2)^2 - 2\xi_1\xi_2\} \\ &= 4n\mu_2 + n\mu_3^2/\mu_2^2, \end{aligned} \quad (10)$$

$$(x_1-x_2)^2 = (\xi_1+\xi_2)^2 - 4\xi_1\xi_2 = 4\mu_2 + \mu_3^2/\mu_2^2, \quad (11)$$

and the sum of the weights is the ratio of these two expressions, as we see from the first term in (2). This gives a useful check on the arithmetic.

In practice it is not necessary to use the exact values of  $x_1$  and  $x_2$ . Approximations to them will suffice to make the correlation between the errors negligible, and the curvature, in any case small in the type of problem considered, will make a negligible contribution. The most convenient method of solution will usually be to solve by fitting a linear function as usual and to find  $y_1$  and  $y_2$  and their uncertainties by the usual method. If desired we can use  $\chi^2$  to test the fit at other values, and if there is a clear departure from the linear form we may either estimate a curvature term or use shorter intervals. The latter course is the more convenient, since the curvature if genuine can be found more accurately later by comparing different ranges.

In practice it is convenient to begin by referring all values of  $x$  to an arbitrary zero near the middle of the range. Then the normal equations to find a linear form

$$y = a + bx \quad (12)$$

will be

$$na + b \sum wx = \sum wy, \quad (13)$$

$$a \sum wx + b \sum wx^2 = \sum wxy; \quad (14)$$

and the second, after eliminating  $a$ , gives

$$b\{\sum wx^2 - (\sum wx)^2/n\} = \sum wxy - \bar{x} \sum wy. \quad (15)$$

The coefficient of  $b$  is

$$\sum w(\xi + \bar{x})^2 - n\bar{x}^2 = \sum w\xi^2 = n\mu_3, \quad (16)$$

so that  $\mu_3$  is found by simple division in the ordinary course of a least squares solution. If we write

$$\sum wx^3 = n\lambda_3, \quad (17)$$

$$\begin{aligned} \text{we have} \quad n\lambda_3 &= \sum w(\xi + \bar{x})^3 = n\mu_3 + 3n\mu_2\bar{x} + n\bar{x}^3, & (18) \\ \text{and therefore} \quad \mu_3 &= \lambda_3 - 3\mu_2\bar{x} - \bar{x}^3. & (19) \end{aligned}$$

The solution is easy and, even if the function is capable of being represented by a polynomial, nearly the whole of the original information is preserved in the summary values. These will not in general be equally spaced, but interpolation can then be done by divided differences.† The method has been extensively used in seismology, where the original intervals and weights were usually unequal. With this method this introduces no difficulty. One feature was found here that may have further application. The curvature terms are rather large, but the higher ones small. For both  $P$  and  $S$  waves the times of transmission were found to be fitted from about  $20^\circ$  to  $90^\circ$  by quadratics, within about  $1/150$  of the whole range of variation, though inspection of the small residuals against them showed that these were systematic. Convenient quadratics were therefore subtracted from the observed times, and linear forms were fitted to the departures from these for the separate ranges. Summary values were found at distances rounded to the nearest multiple of  $0.5^\circ$ , and added to the quadratics at these distances, and finally the whole was interpolated to  $1^\circ$ . There was no previous reason why quadratics should give so good a fit, but the fact that they did made further smoothing easier.‡

The choice of ranges to summarize is mainly a matter of convenience. The only condition of importance is that they must not be long enough for a cubic term to become appreciable within them, since its values at  $x_1$  and  $x_2$  will not in general vanish. This can be tested afterwards by comparing the divided differences of the summary values with their uncertainties. If the third differences are found significant it may be worth while to use shorter ranges; if not, we may get greater accuracy by taking longer ones.

A solution has been found for the problem of finding three summary values from a quadratic determined by least squares, such that their uncertainties are independent of one another and their values unaffected by a possible cubic term.§ It has not, however, been found so far to give enough improvement to compensate for the increased complication in the arithmetic.

#### 4.6. Correction of a correlation coefficient. In a common class of

† Whittaker and Robinson, *Calculus of Observations*, ch. ii; H. and B. S. Jeffreys, *Methods of Mathematical Physics*, 237–41.

‡ *M.N.R.A.S. Geophys. Suppl.* 4, 1937, 172–9, 239–40.

§ *Proc. Camb. Phil. Soc.* 33, 1937, 444–50.

problem the observations as actually recorded are affected by errors that affect the two variables independently, and whose general magnitude is known from other sources. They may be errors of observation, and it is a legitimate question to ask what the correlation would be if the observations were made more accurate. The observations may have been grouped, and we may ask what the correlation would be if the original data were available. We represent these additional sources of error by standard errors  $\sigma_0$ ,  $\xi_0$ , and continue to use  $\sigma$  and  $\tau$  for the ideal observations of which the available ones are somewhat imperfect modifications. But now the expectations of  $x^2$ ,  $y^2$ , and  $xy$  will be  $\sigma^2 + \sigma_0^2$ ,  $\tau^2 + \tau_0^2$ ,  $\rho\sigma\tau$ , since the contributions of the additional error to  $x$  and  $y$  are independent. A normal correlation surface corresponding to these expectations will still represent the conditions of observation if the additional error is continuous. If it is due to grouping we can still use it as a convenient approximation. But for this surface the proper scale parameters and correlation coefficient will be

$$\sigma' = (\sigma^2 + \sigma_0^2)^{1/2}, \quad \tau' = (\tau^2 + \tau_0^2)^{1/2}, \quad \rho' = \rho\sigma\tau/\sigma'\tau'. \quad (1)$$

Now we have seen for one unknown that the best treatment of a known component of the standard error is to continue to use the  $d\sigma/\sigma$  rule for the prior probability of the *whole* standard error, merely truncating it so as to exclude values less than the known component. Consequently the analysis for the estimation of the correlation coefficient stands with the substitution of accented letters as far as 3.8 (10). Thus

$$P(d\rho \mid \theta, H) \propto \frac{(1 - \rho'^2)^{1/2n}}{(1 - \rho'^2)^{n-1/2}} d\rho. \quad (2)$$

If then  $\sigma_0$  and  $\tau_0$  are small compared with  $\sigma$  and  $\tau$ , it will be possible, within the range of probable values of the parameters, to take the prior probabilities of  $\rho$  and  $\rho'$  proportional; and then we can apply the  $(z, \zeta')$  transformation to  $r$  and  $\rho'$  as before. The result may be written

$$\zeta' = z - \frac{5r}{2n} \pm \frac{1}{\sqrt{(n-1)}}, \quad (3)$$

from which the probability distribution of  $\rho'$  follows at once. To derive that of  $\rho$  we must multiply all values of  $\rho'$  by the estimate of  $\sigma'\tau'/\sigma\tau$ , which will be

$$\mu = \frac{st}{(s^2 - \sigma_0^2)^{1/2}(t^2 - \tau_0^2)^{1/2}}. \quad (4)$$

The procedure is thus simply to multiply the correlation and its uncertainty, found as for the standard case, by the product of the ratios of the uncorrected and corrected standard errors in the two



variables. Where the additional variation is due to grouping, this is the product of the ratios without and with Sheppard's corrections.

This device for correcting a correlation coefficient has been derived otherwise from consideration of expectations; but there is a complication when the correlation is high, since it is sometimes found that the 'corrected' correlation exceeds 1. This means that the random variation has given an  $r$  somewhat greater than  $\rho'$ , which is already high, and if the usual correction is applied we are led to an impossible result. The solution is in fact simple, for the only change needed is to remember that the prior probability of  $\rho$  is truncated at  $\pm 1$ . We have therefore only to truncate the posterior probability at  $\rho = \pm 1$  also. If  $\mu r > 1$  the probability density will be greatest at  $\rho = 1$ .

Such treatment is valid for one estimation, but when many have to be combined there is a complication analogous to that for negative parallaxes in astronomy (cf. p. 142). The data must always be combined *before* truncation. To truncate first and then take a mean would lead to systematic underestimates of correlation.

**4.7. Rank correlation.** This method, introduced by Spearman and modified by Pearson, is extensively used in problems where a set of individuals are compared in respect of two properties, which either are not measurable or whose measures do not follow the normal law even roughly. The chief applications are in psychology, where there are few definite standards of measurement, but it is possible to arrange individuals in orders with respect to two or more abilities. Then the orders can be compared without further reference to whether the abilities have received any quantitative measure at all, or if they have, whether this measure follows a normal law of chance. It is clear that if one ability is a monotonic function of the other, no matter how the measures may be made, the orders will either be the same or exactly opposite, so that the amount of correspondence between the orders will indicate the relation, if any, between the abilities. Spearman's proposal, then, was to assign numbers 1 to  $n$  to the observed individuals in respect of each ability, and then to consider the differences between their placings. If  $x$  and  $y$  are the placings of the same individual, the coefficient  $R$  was defined† by

$$R = 1 - \frac{3 \sum |x-y|}{n^2-1}. \quad (1)$$

This coefficient has a peculiarity: If the orders are the same, we have

† *Brit. Journ. Psych.* 2, 1906, 89-108.

$\sum |x-y| = 0$ , and  $R = 1$ . But if they are opposite we have, for four members,

$x$	$y$	$ x-y $
1	4	3
2	3	1
3	2	1
4	1	3
		$\frac{8}{8}$

and 
$$R = 1 - \frac{3 \times 8}{15} = -0.6.$$

Thus complete reversal of the order does not simply reverse the sign of  $R$ . This formula has been largely superseded by another procedure also mentioned by Spearman, namely that we should simply work out the correlation coefficient between the placings as they stand. The mean being  $\frac{1}{2}(n+1)$  in each case, this will be

$$r = \frac{\sum \{x - \frac{1}{2}(n+1)\} \{y - \frac{1}{2}(n+1)\}}{\sqrt{[\sum \{x - \frac{1}{2}(n+1)\}^2 \sum \{y - \frac{1}{2}(n+1)\}^2]}} \quad (2)$$

which can also be written

$$r = 1 - \frac{6 \sum (x-y)^2}{n^3 - n}. \quad (3)$$

This expression is known as the rank correlation coefficient. It is  $+1$  if the orders are the same and  $-1$  if they are opposite.

The formula needs some modification where some individuals in either series are placed equal. A formula for the correction is given by 'Student'† but it is possibly as easy to work out  $r$  directly, giving the tied members the mean number that they would have if the tie were separated.

The rank correlation, while certainly useful in practice, is difficult to interpret. It is an estimate, but what is it an estimate of? That is, it is calculated from the observations, but a function of the observations has no relevance beyond the observations unless it is an estimate of a parameter in some law. Now what can this law be? For  $r = 1$  and  $r = -1$  the answer is easy; the law is that each ability is a monotonic function of the other. If the abilities are independent, again, the expectation of  $r$  is 0, and if  $r$  is found 0 in an investigation it will naturally be interpreted as an indication of independence. But for intermediate values of  $r$  the interpretation is not clear. The form (2) itself is the one derived for normal correlation; but the normal correlation

† *Biometrika*, 13, 1921, 263-82.

surface has a maximum in the centre and an infinite range of possible values in all directions. In a given experiment any combination of these might occur. But  $x$  and  $y$  have a finite range of possible values, each of which they can take once and only once. The validity of the form (2) in relation to  $x$  and  $y$  therefore needs further examination.  $r$  may be an estimate of some parameter in a law, but it is not clear what this law can be, and whether  $r$  will be the best estimate for the parameter.

To illustrate the point, suppose that a pair of large samples from different classes have been compared. A pair of smaller samples is taken from them at random. What is the probability distribution of  $r$  for the comparison of these small samples? Except for some extreme cases, nobody knows; but we should want to know whether it depends only on the value of  $r$  for the comparison of the large classes, or whether it depends also on finer features of the relative distribution. In the latter case, if we had only the small samples,  $r$  found from them will not be a sufficient statistic for  $r$  in the large samples.

Pearson† has investigated the relation of  $r$  to normal correlation. If we consider the two laws

$$P(dxdy | \sigma_1, \sigma_2, H) = \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{x^2}{2\sigma_1^2} - \frac{y^2}{2\sigma_2^2}\right) dxdy, \quad (4)$$

$$P(dxdy | \sigma_1, \sigma_2, \rho, H) \\ = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{x^2}{\sigma_1^2} - \frac{2\rho xy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2}\right)\right\} dxdy, \quad (5)$$

both give the same total chance of  $x$  or of  $y$  separately being in a given range. Consequently we can introduce two functions (called by Pearson the *grades*)

$$X = \int_{-\infty}^x \frac{1}{\sqrt{(2\pi)\sigma_1}} \exp\left(-\frac{x^2}{2\sigma_1^2}\right) dx, \quad Y = \int_{-\infty}^y \frac{1}{\sqrt{(2\pi)\sigma_2}} \exp\left(-\frac{y^2}{2\sigma_2^2}\right) dy \quad (6)$$

and eliminate  $x$  and  $y$  in favour of  $X$  and  $Y$ . Then the right of (4) is simply  $dXdY$  for  $X$  and  $Y$  between 0 and 1. Then (5) expressed in terms of  $X$  and  $Y$  gives a distribution within a square, and showing correlation between  $X$  and  $Y$ . Further, such a transformation would be possible for any law of chance; we simply need to take as new variables the chances that  $x$  and  $y$  separately are less than given values. The result will not be a normal correlation surface in either case, and there appears to be no reason to suppose that it would always be of the

† *Drapers' Co. Research Mems., Biometric Series*, 4, 1907, 1-39.

same functional form. Nevertheless, one property of normal correlation will persist. The exponent in (5) can be written

$$-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{x}{\sigma_1}-\frac{\rho y}{\sigma_2}\right)^2+(1-\rho^2)\frac{y^2}{\sigma_2^2}\right\} \quad (7)$$

and we can take  $x' = x - \rho\sigma_1 y/\sigma_2$  and  $y$  as new variables. These will have independent chances, and then if  $\rho$  tends to 1 the standard error of  $x'$  will tend to 0 and that of  $y$  to  $\sigma_2$ . Thus in the limiting case the normal correlation surface reduces to a concentration along a line and  $y$  is strictly a monotonic function of  $x$ . Analogous relations hold if  $\rho$  tends to  $-1$ . But then  $X$  and  $Y$  will be equal, since  $x$  and  $y$  are proportional.

An analogous transformation applied to any other law will make  $X$  and  $Y$  equal if  $x$  and  $y$  are monotonic functions of each other, not necessarily linear, and  $r$  will be  $+1$  or  $-1$ . Now it seems to me the chief merit of the method of ranks that it eliminates departure from linearity, and with it a large part of the uncertainty arising from the fact that we do not know any form of the law connecting  $x$  and  $y$ . For any law we could define  $X$  and  $Y$ , and then a new  $x$  and  $y$  in terms of them by (6). The result, expressed in terms of these, need not be a normal correlation surface, but the chief difference will be the one that is removed by reference to orders instead of measures.

Accordingly it appears that if an estimate of the correlation, based entirely on orders, can be made for normal correlation, it may be expected to have validity for other laws; the same type of validity as the median of a series of observations has in estimating the median of the law, that is, not necessarily the best that can ever be done, but the best that can be done until we know more about the form of the law itself. But whereas for normal correlation it will estimate departure from linearity, for the more general law it will estimate how far one variable departs from being a monotonic function of the other.

Pearson investigates the expectations of Spearman's two coefficients for large samples of given size derived from a normal correlation surface, and gets

$$E(r) = \frac{6}{\pi} \sin^{-1} \frac{1}{2} \rho$$

so that (8)

$$\rho = 2 \sin(\frac{1}{3} \pi r)$$

is an estimate of  $\rho$  involving only orders. In terms of  $R$  he gets

$$\rho = 2 \cos \frac{1}{3} \pi (1 - R) - 1. \quad (9)$$

The latter has the larger uncertainty. Little further attention has

therefore been paid to  $R$ . The expectation of the square of the random variation in  $r$  leads to the result that, if  $\rho$  is given, the standard error of an estimate of  $\rho$  from  $r$  would be

$$1.0472 \frac{1-\rho^2}{\sqrt{n}} (1 + 0.042\rho^2 + 0.008\rho^4 + 0.002\rho^6). \quad (10)$$

The corresponding formula for a correlation found directly from the measures is  $(1-\rho^2)/\sqrt{n}$ , so that even for normal correlation  $r$  gives a very efficient estimate. Pearson comments on the fact that in some cases where the distribution is far from normal the value of  $\rho$  found from  $r$  is noticeably higher than that found from the usual formula, and seems to think that the fault lies with  $r$ . But if  $x$  was any monotonic function of  $y$  other than a linear one, the usual formula would give  $\rho$  less than 1, whereas the derivation from  $r$  would be 1. Thus if  $y = x^3$  for  $-1 < x < 1$ , we have

$$E(x^2) = \frac{1}{3}, \quad E(x^6) = \frac{1}{7}, \quad E(x \cdot x^3) = \frac{1}{5};$$

$$\rho = \frac{\frac{1}{5}}{(\frac{1}{3} \cdot \frac{1}{7})^{1/2}} = \left(\frac{21}{25}\right)^{1/2} = +0.917.$$

The ranks method puts  $x$  and  $x^3$  in the same order and leads to  $\rho = 1$ ; but that is not a defect of the method, because it does not measure departure from linearity but from monotonicity, and in its proper sense it gives the right answer. The formula based on  $\sum xy$  measures departure from linearity, and there is no inconsistency. Further, there is no reason to suppose that with great departures from normality this formula gives an estimate of anything particular.

Pearson is very critical of Spearman in parts of this paper, but I think that he provides a very satisfactory justification of his coefficient. Spearman has replied,<sup>†</sup> but does not mention the last point, which I think is the chief merit of his method. The rank correlation leads to nearly as good an estimate as the product moment in the case where the latter is definitely the best estimate. It is also right in cases of complete association where  $y$  is a monotonic but not a linear function of  $x$ . In such cases the normal law and normal correlation do not hold, and the product moment would suggest imperfect association between  $x$  and  $y$ . It is also right in testing absence of association. For general use where the law is unknown and may be far from normal it seems in this respect to be definitely better than  $\overline{xy}/s_1 s_2$ . Its defect is that we still have not succeeded in stating just what it measures in general. The normal correlation surface is a complete statement of the joint chance of two

<sup>†</sup> *Brit. Journ. Psych.* 3, 1910, 271-95. See also Yule, *J. R. Stat. Soc.* 70, 1907, 656.

variables, and  $\rho$  is a parameter in this law. The extension to non-normal correlation would still require such a law, containing one new parameter, leading to an expression for the joint chance of  $n$  individuals being arranged in any two orders with respect to two abilities, and stated entirely in terms of those orders. Such a law has not been found; I have searched for possible forms, but all have been either intrinsically unsatisfactory in some respect or led to mathematical difficulties that I, at any rate, have not succeeded in overcoming. Till this is done there will be some doubt as to just what we mean quantitatively, in regard to two quantities both subject to a certain amount of random variation, by the amount of departure from monotonicity. Should the law involve  $\exp\{-\alpha |X - Y|\}$  or  $\exp\{-\alpha(X - Y)^2\}$ , for instance, we should be led to different functions of the observed positions to express the best value of  $\alpha$ ; and to decide between them would apparently need extensive study of observations similar to those used to test whether the normal law of errors holds for measures. It cannot be decided *a priori*, and until we have some way of finding it by experiment some indefiniteness is inevitable.

Pearson's formula for the standard error of the correlation coefficient, as found for the normal correlation surface by the method of ranks, does not give the actual form of the probability distribution, which is far from normal unless the number of observations is very large. But his estimates of uncertainty for the correlation coefficient found by the most efficient method in this case, and for that found from the rank coefficient, have been found by comparable methods, and two functions with the same maximum, the same termini at  $\pm 1$ , and the same second moment about the maximum, are unlikely to differ greatly. It appears therefore that we can adapt the formulae 3.8 (25) and (26) by simply multiplying the standard error of  $\zeta$  by

$$1.0472(1 + 0.042\rho^2 + 0.008\rho^4 + 0.002\rho^6)$$

for the estimated  $\rho$ .

An alternative method is given by Fisher and Yates. One disadvantage of the correlation between ranks as they stand is that if we have, say, 10 pairs, in the same order, the effect of interchanging members 1 and 2 in one set is the same as that of interchanging members 5 and 6. That is, the correlations of the series

1, 2, 3, 4, 5, 6, 7, 8, 9, 10

with

2, 1, 3, 4, 5, 6, 7, 8, 9, 10

and with

1, 2, 3, 4, 6, 5, 7, 8, 9, 10

are the same. But if the series are the results of applying ranking to a

normal correlation surface this is wrong, for the difference between members 1 and 2 would ordinarily be much larger than that between members 5 and 6. Fisher and Yates† deal with this by using the ranks, as far as possible, to reconstruct the measures that would be obtained in a normal correlation. If as before we use  $X$  to denote the chance of an observation less than  $x$ , where  $x$  is derived from the normal law with  $\sigma = 1$ , the chance of  $p-1$  observations being less than  $x$ ,  $n-p$  greater than  $x+dx$ , and one between  $x$  and  $x+dx$ , is

$$\frac{n!}{(p-1)!(n-p)!} X^{p-1}(1-X)^{n-p} dX,$$

and this is the chance that the  $p$ th observation will lie in a range  $dx$ . The expectation of  $x$  for the  $p$ th observation in order of rank is therefore

$$x_p = \frac{n!}{(p-1)!(n-p)!} \int_0^1 X^{p-1}(1-X)^{n-p} x dX,$$

and if this is substituted for the rank we have a variable that can be used to find the correlation coefficient directly without transformation. This avoids the above difficulty. It makes the expectation of the sum of the squares of the differences between the actual measures and the corresponding  $x_p$  a minimum. Fisher and Yates give a table of the suitable values of  $x_p$  for  $n$  up to 30. The uncertainty given by this method must be larger than that for normal correlation when the data are the actual measures, and smaller than for the correlation derived from Spearman's coefficient, and the difference is not large. Fisher and Yates tabulate  $\sum x_p^2$ , and Fisher tells me privately that the allowance would be got by multiplying the uncertainty by  $(n/\sum x_p^2)^{1/2}$ , but the proof has not been published.

The difference between Pearson's method and Fisher's recalls a similar problem for one variable (cf. 3.61). Re-scaling may obscure an essential feature of the distribution, and presumably will also do so for distributions for two variables. I think that what is needed is rather some method of analysis, like the use of the median in 4.4, such that the results will be as insensitive as possible to the actual form of the law; completely insensitive they cannot be.

A further way of estimating rank correlation is given by Kendall.‡

**4.71. Grades and Contingency.** The method of ranks can be extended to a contingency table classified by rows and columns. Pearson's

† *Statistical Tables*, 1938, pp. 13, 50-1.

‡ *The Advanced Theory of Statistics*, ch. 16, especially pp. 391-4, 403-8.

analysis actually leads to (8) and (10) by a consideration of the correlation between grades, which are the quantities I have denoted by  $X$  and  $Y$  and are called  $g_1$  and  $g_2$  by him. If the quantities correlated are magnitudes and we have a series of measures, then for the normal correlation surface  $X$  and  $Y$  will be read from a table of the error function and known for each observation with the same order of accuracy as the measures. Then the rank correlation will be the correlation between  $X$  and  $Y$ . If we have the orders of individuals with regard to two properties, these provide the estimated  $X$  and  $Y$ , from which we can calculate the rank correlation and proceed to  $\rho$ , in possibly an extended sense if the correlation is not normal. When data have been classified the same will hold approximately, on account of the small effect of even rather drastic grouping on the estimates. The following table of the relation of colours and spectral types of stars provides an example.† The spectral types are denoted by  $x$ , the colours by  $y$ , as follows.‡

$x$		$y$
1	Helium stars	1 White
2	Hydrogen stars	2 White with faint tinge of colour
3	$\alpha$ Carinae type	3 Very pale yellow
4	Solar stars	4 Pale yellow
5	Arcturus type	5 Full yellow
6	Aldebaran type	6 Ruddy
7	Botelgeuse type	

$x$	$y$	1	2	3	4	5	6	Total	Mean rank $X$ 100 ×
1		125	146	8	3	0	0	282	-5.9
2		168	195	14	0	0	0	377	-2.6
3		3	97	23	8	6	0	137	0
4		0	41	77	33	29	0	180	+1.6
5		0	15	86	77	63	0	241	+2.8
6		0	0	4	22	43	6	75	+4.4
7		0	3	2	39	19	5	68	+5.1
Total		296	497	214	182	160	11	1,360	
Mean rank $Y$ 100 ×		-7.5	-3.6	0	+2.0	+3.7	+4.5		

For convenience the zero of rank is taken in the middle of the third group for both  $X$  and  $Y$ , and the ranks given are the means of the placings relative to this zero, and divided by 100. Then we find

$$\begin{aligned}\sum X &= -1004, & \sum Y &= -3003, \\ \sum X^2 &= 17939, & \sum Y^2 &= 26233, & \sum XY &= +15914.\end{aligned}$$

† W. S. Franks, *M.N.R.A.S.* 67, 1907, 539-42. Quoted by Brunt, *Combination of Observations*, p. 170.

‡ What Franks calls a white star would be called bluish by many observers, who would call his second class white.



The mean ranks are therefore at  $X = -0.7$ ,  $Y = -2.2$ ; to reduce to the means we must apply to  $\sum X^2$ ,  $\sum Y^2$ , and  $\sum XY$  the corrections  $-1004 \times 0.7$ ,  $-3003 \times 2.2$ ,  $-1004 \times 2.2$ . Also we must correct  $\sum X^2$  and  $\sum Y^2$  for grouping. In the first row for  $x$ , for instance, grouping has made a contribution of  $\frac{1}{12} 282(2.82)^2$  to  $\sum X^2$ , and so on. It does not affect the product systematically. Allowing for this we should reduce  $\sum X^2$  and  $\sum Y^2$  by a further 826 and 1405. Thus the corrected values are

$$\sum X^2 = 16410; \quad \sum Y^2 = 18192; \quad \sum XY = +13765.$$

These give  $r = +0.798$ .

To convert to an analogue of the correlation coefficient we must take

$$\rho = 2 \sin(0.5236 \times 0.798) = 0.812.$$

Applying the  $z$  transformation we get

$$\zeta = 1.133 - 0.003 \pm 0.037.$$

This uncertainty is a little too low, since it has allowed for grouping, which should not be done in estimating uncertainties. This has altered both  $X^2$  and  $Y^2$  by about 5 per cent., and we should increase the standard error of  $\zeta$  by the same amount. Also we should multiply by 4.7 (10) because we are working with ranks and not measures. This is 1.09. Hence (ranges corresponding to the standard error)

$$\zeta = 1.130 \pm 0.042 = 1.088 \text{ to } 1.172,$$

$$\rho = +0.796 \text{ to } +0.825.$$

Brunt, from the above data, using Pearson's coefficient of mean square contingency, gets  $\rho = +0.71$ . The difference is presumably due to the skewness of the distribution, the greatest concentration being in one corner of the table. I think that my larger value gives a better idea of the closeness of the correspondence. But I think that the use of this coefficient to estimate association is undesirable for other reasons. In a rectangular contingency table  $\chi^2$  may be computed against the hypothesis of proportionality of the chances in the rows, and Pearson defines the mean square contingency by

$$\phi^2 = \chi^2/N,$$

where  $N$  is the whole number of observations.† He then considers the laws for correlations 0 and  $\rho$ , on the former of which proportionality would hold, and works out, against the chances given by it, the value of  $\phi^2$  supposing the number of observations very large and distributed

† *Drapers' Co. Res. Memos., Biometric Series, 1, 1904.*

exactly in proportion to the expectations given by normal correlation  $\rho$ . The result for this limiting case is  $\rho^2/(1-\rho^2)$ ; and hence

$$\rho^2 = \frac{\phi^2}{1+\phi^2}$$

is suggested as a possible means of estimating  $\rho$ . Unfortunately in practice we are not dealing with limiting cases but with a finite number of observations classified into groups, and even if the two variables were strictly independent the sampling errors would in general make  $\chi^2$  about  $(m-1)(n-1)$ , where  $m$  and  $n$  are the numbers of rows and columns. For an actual series of observations  $\phi^2$  will always be positive, and  $r$  will be estimated by this method as about  $\{(m-1)(n-1)/N\}^{1/2}$  if the variations are independent. This is not negligible. But also if there are any departures from proportionality of the chances whatever, irrespective of whether they are in accordance with a normal correlation, they will contribute to  $\chi^2$  and therefore to the estimate of  $\rho^2$ . The excess chances might, for instance, be distributed alternately by rows and columns so as to produce a chessboard pattern; this is nothing like correlation, but the  $\phi^2$  method would interpret it as such. Or there might be a failure of independence of the events, leading to a tendency for several together to come into the same compartment; an extension of the idea that we have had in the negative binomial distribution. This would not affect the distribution of the expectation, but it would increase  $\phi^2$ . On the other hand, grouping will reduce  $\phi^2$  if the correlation is high. Accordingly I think that this function, or any other function of  $\chi^2$ , should be used as an estimate only when the only parameter considered is one expressing intraclass correlation or non-independence of the events. It is not suited to estimate the normal correlation coefficient because too many other complications can contribute to it and produce a bias. In the above case, however, the departure from normality itself has led to a greater effect in the opposite direction, and in the circumstances it seems that this way of estimating association would be best abandoned.

**4.8. The estimation of an unknown and unrestricted integer.** The following problem was suggested to me several years ago by Professor M. H. A. Newman. A man travelling in a foreign country has to change trains at a junction, and goes into the town, of the existence of which he has only just heard. He has no idea of its size. The first thing that he sees is a tramcar numbered 100. What can he infer about the number of tramcars in the town? It may be assumed for the purpose that they are numbered consecutively from 1 upwards.

The novelty of the problem is that the quantity to be estimated is a positive integer, with no apparent upper limit to its possible values. A uniform prior probability is therefore out of the question. For a continuous quantity with no upper limit the  $dv/v$  rule is the only satisfactory one, and it appears that, apart from possible complications at the lower limit, we may suppose here that if  $n$  is the unknown number

$$P(n | H) \propto n^{-1} + O(n^{-2}). \quad (1)$$

Then the probability, given  $n$ , that the first specimen will be number  $m$  in the series is

$$P(m | n, H) = 1/n \quad (m \leq n) \quad (2)$$

and therefore  $P(n | m, H) \propto n^{-2} + O(n^{-3}) \quad (n \geq m).$  (3)

If  $m$  is fairly large the probability that  $n$  exceeds some definite value  $n_0$  will be nearly

$$P(n > n_0 | m, H) = \sum_{n_0+1}^{\infty} n^{-2} / \sum_m^{\infty} n^{-2} = \frac{m}{n_0}, \quad (4)$$

nearly. With one observation there is a probability of about  $\frac{1}{2}$  that  $n$  is not more than  $2m$ .

I have been asked this question several times and think that an approximate solution may be worth recording. The interesting thing is that the questioners usually express a feeling that there is something special about the value  $2m$ , without being able to say precisely what it is. The adopted prior probability makes it possible to say how a single observation can lead to intelligible information about  $n$ , and it seems to be agreed that it would do so. I see no way, however, of fixing the terms of order  $n^{-2}$ .

The extension to the case where several members of the series are observed is simple, and is closely analogous to the problem of finding a rectangular distribution from a set of measures.

**4.9. Artificial randomization.** This technique in experimental design has been greatly developed by Fisher,<sup>†</sup> and more recently by Yates,<sup>‡</sup> chiefly in relation to agricultural experiments. The primary problem in the work is to compare the productivities of different varieties of a plant and the effects of different fertilizers and combinations of fertilizers. The difficulty is that even if the same variety is planted in a number of plots and all receive the same treatment, the yields differ. Such tests are called uniformity trials. This would not affect the work

<sup>†</sup> *The Design of Experiments*, 1935.

<sup>‡</sup> *J. R. Stat. Soc. Suppl.* 2, 1935, 181-223; *The Design and Analysis of Factorial Experiments*, Imp. Bur. of Soil Science, 1937.

if the yields were random; if they were, the plot yields could be taken as equations of condition for the varietal and treatment differences and the solution completed by least squares, thus obtaining the best possible estimates and a valid uncertainty. Unfortunately they are not random. In uniformity trials it is habitually found that there is a significant gradient in the yield in one or other direction on the ground. Even when this is estimated and taken into account it is found that there is a marked positive correlation between neighbouring plots. Further, many fields have at some stage of their history been laid out for drainage into a series of parallel ridges and furrows, which may leave a record of themselves in a harmonic variation of fertility. The result is that the analysis of the variation of the plot yields into varietal and treatment differences and random error does not represent the known facts; these ground effects must be taken into account in some way. The best way, if we want to get the maximum accuracy, would be to introduce them explicitly as unknowns, form normal equations for them also, and solve. Since the arrangement of the plots is at the experimenter's disposal, his best plan is to make it so that the equations for the various unknowns will be orthogonal. One of the best ways of doing this is by means of the Latin square. If the plots are arranged in a  $5 \times 5$  square to test five varieties, and each variety occurs just once in each row and each column, the estimates of the differences between the varieties will be the differences of the means of the plots containing them, irrespective of the row and column differences of fertility. But unfortunately the correlation between neighbouring plots still prevents the outstanding variation from being completely random. If it was, all Latin squares would be equally useful. But suppose that we take Cartesian coordinates of position at the centre of each square, the axes being parallel to the sides. Then if variations of fertility are completely expressed by the row and column totals they are expressible in the form

$$F = a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4 + b_1y + b_2y^2 + b_3y^3 + b_4y^4.$$

For with suitable choices of the  $a$ 's and  $b$ 's it will be possible to fit all the row and column totals exactly. But this contains no product terms, such as  $xy$ . In certain conditions this might be serious; for if  $x^2$  and  $y^2$  produce a significant variation it would only be for one special orientation of the sides that the  $xy$  term would be absent, and if the plots containing one variety all correspond to positive  $xy$  and all containing another to negative  $xy$ , part of the difference between the means for these sets of plots will be due to the  $xy$  term in the fertility and not to

the differences of the varieties. This will happen with the most obvious design, namely

<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>
<i>E</i>	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>
<i>D</i>	<i>E</i>	<i>A</i>	<i>B</i>	<i>C</i>
<i>C</i>	<i>D</i>	<i>E</i>	<i>A</i>	<i>B</i>
<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>A</i>

Here varieties *C* and *D* have positive or zero  $xy$  every time, while *A* has negative or zero  $xy$  every time. If, then, the  $x^2$  and  $y^2$  terms should be eliminated, should we not estimate and eliminate  $xy$  too? On the face of it it will usually be more important than higher terms such as  $x^4$ ; but the real question is, where are we to stop? If we should keep the whole of the terms up to the fourth power, we shall need to eliminate 6 extra terms, leaving only 6 to give an estimate of the random variation; if we should go to  $x^4y^4$  we should be left with no information at all to separate varieties from fertility. We must stop somewhere, and for practical reasons Fisher introduces at this stage another method of dealing with  $xy$ , which leaves it possible to use the plot means alone to estimate the varietal differences and at the same time to treat the outstanding variation as if it were random, though in fact it is not. Possibly it is often an unnecessary refinement to eliminate the higher terms completely, as he does, but the analysis doing so is easier than it would be to omit them and find the lower ones by least squares, and it does no harm provided sufficient information is left to provide a good estimate of the uncertainty. But there might be a serious danger from  $xy$ . In a single  $5 \times 5$  square each variety occurs only 5 times, and some of this information, effectively 1.8 plots per variety, is sacrificed in eliminating the row and column fertility effects. But if we use the usual rules for estimating uncertainty they will suppose that when we have allowed for rows, columns, and varieties, the rest of the variation is random. If there is an  $xy$  term, this will be untrue, since the sign of this term in one plot will determine that in every other. With some arrangements of the varieties the contributions to the means of the plots with the same variety due to  $xy$  will be more, with others less, than would be expected if they were completely random contributions with the same mean square. Consequently it will not be valid to treat the outstanding variation as random in estimating the uncertainty of the differences between the varieties.  $xy$  could be introduced explicitly,

with an unknown coefficient to be found from the data, and then on eliminating it the results would be unaffected by it. But this would mean appreciable increase of labour of computation, and the possibility of still higher terms might then have to be considered.

Again, it is usual to lay out two or three squares to reduce the uncertainty. If the same design was used for three squares there would be a  $\frac{1}{4}$  chance that every variety would have  $\sum xy$  for its plots with the same sign in every square. This is not a negligible chance; and though the differences of the  $\sum xy$  for the varieties in one square might be unimportant, their contribution to the estimated total differences would be multiplied by 3 in three squares, while their contribution to the estimated standard error of these totals, assuming randomness, would only be multiplied by  $\sqrt{3}$ . Thus if the design is simply copied, and an  $xy$  term is present, there is an appreciable chance that it may lead to differences that would be wrongly interpreted as varietal.

Fisher proceeds, instead of determining the  $xy$  term, to *make it* into a random error. This is done by arranging the rows and columns of every square at random. Thus if we start with the arrangement given above, we have in the first column the order *AEDCB*. By a process such as card shuffling we rearrange these letters in a new order, such as *CADEB*. The rows are then rearranged, keeping each row intact, so as to bring the letters in the first column into this order. The letters in the first row are now in the order *CDEAB*. Shuffling these we get *ECBAD*; and now rearranging the columns we get the final arrangement

<i>E</i>	<i>C</i>	<i>B</i>	<i>A</i>	<i>D</i>
<i>C</i>	<i>A</i>	<i>E</i>	<i>D</i>	<i>B</i>
<i>A</i>	<i>D</i>	<i>C</i>	<i>B</i>	<i>E</i>
<i>B</i>	<i>E</i>	<i>D</i>	<i>C</i>	<i>A</i>
<i>D</i>	<i>B</i>	<i>A</i>	<i>E</i>	<i>C</i>

The varieties would be laid out in this order in an actual square; but for the second and third squares entirely separate rearrangements must be made. There is no such thing as an intrinsically random arrangement. The whole point of the design is that if there is an  $xy$  term in the fertility, its contribution to any varietal total in one square shall give no information relevant to the total in another square. Card shuffling is fairly satisfactory for this purpose because one deal does give little or no information relevant to the next. But if the deal is simply

copied the terms in  $xy$  for one square will give information about their values in the others, and the shuffling fails in its object. An arrangement can only be random once.

This procedure, highly successful in practice, shows well the conditions for the use of artificial randomization. In the first place, the square is not randomized completely. The rule that each variety shall occur just once in every row and in every column is absolute. If 25 cards were lettered, 5 with  $A$ , 5 with  $B$ , and so on, and shuffled, the result would be that some letters would be completely absent from some columns and appear two or three times in others. The result would be a loss of accuracy in the estimation of the linear gradients, which could therefore not be allowed for with so much accuracy, and this would increase the final uncertainty of the varietal differences. Here is the first principle: we must not try to randomize a systematic effect that is known to be considerable in relation with what we are trying to find. The design must be such that such effects can be estimated and eliminated as accurately as possible, and this is done best if we make an error in an unknown of either set contribute equally to the estimates of all unknowns of the other sets. But this condition imposes a high degree of system on the design, and any attempt at randomness must be within the limits imposed by this system. In some discussions there seems to be a confusion between the design itself and the method of analysing the results. The latter is always to take the means of the plot yields with the same variety to give the estimates of the varietal differences. It is not asserted that this is the best method. If the  $xy$  term was allowed for explicitly the analysis would, in general, be more complicated, but elimination of the variation due to it would leave results of a higher accuracy, which would not, however, rest simply on the differences of the means. The method of analysis deliberately sacrifices some accuracy in estimation for the sake of convenience in analysis. The question is whether this loss is enough to matter, and we are considering again the efficiency of an estimate. But this must be considered in relation to the purpose of the experiment in the first place. There will in general be varietal differences; we have to decide whether they are large enough to interest a farmer, who would not go to the expense of changing his methods unless there was a fairly substantial gain in prospect. There is, therefore, a minimum difference that is worth asserting. It is, however, also important that differences asserted should have the right sign, and therefore the uncertainty stated by the method must be substantially less than the minimum difference that would interest the farmer.

So long as this condition is satisfied it is not important whether the probability that the difference has the wrong sign is 0.01 or 0.001. The design and the method of analysis are therefore, for this purpose, combined legitimately, provided that together they yield an uncertainty small enough for interesting effects not to be hidden by ground effects irrelevant to other fields and deliberately ignored. Previous experiments have usually indicated the order of magnitude of the uncertainty to be expected, with a given design, and it is mainly this that determines the size and number of the plots. This information, of course, is vague, and Fisher and Yates are right in treating it as previous ignorance when they have the data for the actual experiment, which are directly relevant. But it has served to suggest what effects are worth eliminating accurately and what can be randomized without the subsequent method of analysis, treating them as random, giving an uncertainty too large for the main objects of the experiment to be fulfilled. In different conditions, however, the effects that should be eliminated and those that may be randomized and henceforth treated as random will not necessarily be the same.†

The same principles arise in a more elementary way in the treatment of rounding-off errors in computation. If an answer is wanted to one decimal, the second decimal is rounded off so that, for instance, 1.87 is entered as 1.9 and 1.52 as 1.5. If the rejected figure is a 5 it is rounded to the nearest even number; thus 1.55 is entered as 1.6 and 1.45 as 1.4. Thus these minor errors are made random by their association with observational error and by the fact that there is no reason to expect them to be correlated with the systematic effects sought. If rounding-off was always upwards or downwards it would produce a cumulative error in the means.

Most physicists, of course, will envy workers in subjects where uninteresting systematic effects can be randomized, and workers dealing with phenomena as they occur in nature will envy those who can design their experiments so that the normal equations will be orthogonal.

† See also 'Student', *Biometrika*, 29, 1938, 363-79; E. S. Pearson and J. Neyman, *ibid.* 29, 1938, 380-8; E. S. Pearson, *ibid.* 30, 1938, 159-79; F. Yates, *ibid.* 30, 1939, 440-66; Jeffreys, *ibid.* 31, 1939, 1-8.



# V

## SIGNIFICANCE TESTS: ONE NEW PARAMETER

'Which way ought I to go to get from here?'

'That depends a good deal on where you want to get to,' said the Cat.

'I don't much care where——' said Alice.

'Then it doesn't matter which way you go,' said the Cat.

LEWIS CARROLL, *Alice in Wonderland*.

**5.0. General discussion.** THE general principles of significance tests have been stated at the beginning of Chapter III. We need only recall that our problem is to compare a specially suggested value of a new parameter, often 0, with the aggregate of other possible values. We do this by enunciating the hypotheses  $q$ , that the parameter has the suggested value, and  $q'$ , that it has some other value to be determined from the observations. We shall call  $q$  the *null hypothesis*, following Fisher, and  $q'$  the *alternative hypothesis*. To say that we have no information initially as to whether the new parameter is needed or not we must take

$$P(q | H) = P(q' | H) = \frac{1}{2}. \quad (1)$$

But  $q'$  involves an adjustable parameter,  $\alpha$  say, and

$$P(q' | H) = \sum P(q', \alpha | H) \quad (2)$$

over all possible values of  $\alpha$ . We take  $\alpha$  to be zero on  $q$ . Let the prior probability of  $d\alpha$ , given  $q'H$ , be  $f(\alpha)d\alpha$ , where

$$\int f(\alpha)d\alpha = 1, \quad (3)$$

integration being over the whole range of possible values when the limits are not given explicitly. Then

$$P(q' d\alpha | H) = \frac{1}{2}f(\alpha)d\alpha. \quad (4)$$

We can now see in general terms that this analysis leads to a significance test for  $\alpha$ . For if the maximum likelihood solution for  $\alpha$  is  $a \pm s$ , the chance of finding  $a$  in a particular range, given  $q$ , is nearly

$$P(da | qH) = \frac{1}{\sqrt{(2\pi)s}} \exp\left(-\frac{a^2}{2s^2}\right) da, \quad (5)$$

and the chance, given  $q'$  and a particular value of  $\alpha$ , is

$$P(da | q' \alpha H) = \frac{1}{\sqrt{(2\pi)s}} \exp\left\{-\frac{(a-\alpha)^2}{2s^2}\right\} da. \quad (6)$$

Hence by the principle of inverse probability

$$P(q | aH) \propto \frac{1}{\sqrt{(2\pi)s}} \exp\left(-\frac{a^2}{2s^2}\right), \quad (7)$$

$$P(q' d\alpha | aH) \propto \frac{1}{\sqrt{(2\pi)s}} f(\alpha) \exp\left\{-\frac{(a-\alpha)^2}{2s^2}\right\} d\alpha. \quad (8)$$

It is to be understood that in pairs of equations of this type the sign of proportionality indicates the same constant factor, which can be adjusted to make the total probability 1.

Consider two extreme cases. There will be a finite interval of  $\alpha$  such that  $\int f(\alpha) d\alpha$  through it is arbitrarily near unity. If  $a$  lies in this range and  $s$  is so large that the exponent in (8) is small over most of this range, we have on integration, approximately,

$$P(q' | aH) = P(q | aH) \propto \frac{1}{\sqrt{(2\pi)s}}. \quad (9)$$

In other words, if the standard error of the maximum likelihood estimate is greater than the range of  $\alpha$  permitted by  $q'$ , the observations do nothing to decide between  $q$  and  $q'$ .

If, however,  $s$  is small, so that the exponent can take large values, and  $f(\alpha)$  is continuous, the integral of (8) will be nearly  $f(a)$ , and

$$\frac{P(q | aH)}{P(q' | aH)} \doteq \frac{1}{\sqrt{(2\pi)s}f(a)} \exp\left(-\frac{a^2}{2s^2}\right). \quad (10)$$

We shall in general write

$$K = \frac{P(q | \theta H)}{P(q' | \theta H)}. \quad (11)$$

If the number of observations,  $n$ , is large,  $s$  is usually small like  $n^{-1/2}$ . Then if  $a = 0$  and  $n$  large,  $K$  will be large of order  $n^{1/2}$ , since  $f(a)$  is independent of  $n$ . Then the observations support  $q$ , that is, they say that the new parameter is probably not needed. But if  $|a|$  is much larger than  $s$  the exponential will be small, and the observations will support the need for the new parameter. For given  $n$ , there will be a critical value of  $a/s$  such that  $K = 1$  and no decision is reached.

The larger the number of observations the stronger the support for  $q$  will be if  $|a| < s$ . This is a satisfactory feature; the more thorough the investigation has been, the more ready we shall be to suppose that if we have failed to find evidence for  $\alpha$  it is because  $\alpha$  is really 0. But it carries with it the consequence that the critical value of  $a/s$  increases with  $n$  (though that of  $\alpha$  of course diminishes); the increase is very slow,

since it depends on  $\sqrt{(\log n)}$ , but it is appreciable. The test does not draw the line at a fixed value of  $a/s$ .

The simplicity postulate therefore leads to significance tests. The difficulty pointed out before (p. 103) about the uniform assessment of the prior probability was that even if  $\alpha$  was 0,  $a$  would usually be different from 0, on account of random error, and to adopt  $a$  as the estimate would be to reject the hypothesis  $\alpha = 0$  even if it was true. We now see how to escape from this dilemma. Small values of  $|a|$  up to some multiple of  $s$  will be taken to support the hypothesis  $\alpha = 0$ , since they would be quite likely to arise on that hypothesis, but larger values support the need to introduce  $\alpha$ . In suitable cases high probabilities may be obtained for either hypothesis. The possibility of getting actual support for the null hypothesis from the observations really comes from the fact that the value of  $\alpha$  indicated by it is unique.  $q'$  indicates only a range of possible values, and if we select the one that happens to fit the observations best we must allow for the fact that it is a selected value. If  $|a|$  is less than  $s$ , this is what we should expect on the hypothesis that  $\alpha$  is 0, but if  $\alpha$  was equally likely to be anywhere in a range of length  $m$  it requires that an event with a probability  $2s/m$  shall have come off. If  $|a|$  is much larger than  $s$ , however,  $a$  would be a very unlikely value to occur if  $\alpha$  was 0, but no more unlikely than any other if  $\alpha$  was not 0. In each case we adopt the less remarkable coincidence.

This approximate argument shows the general nature of the significance tests based on the simplicity postulate. The essential feature is that we express ignorance of whether the new parameter is needed by taking half the prior probability for it as concentrated in the value indicated by the null hypothesis, and distributing the other half over the range possible.

The above argument contemplates a law  $q$  containing no adjustable parameter and a law  $q'$  containing precisely one. In practice we usually meet one or more of the following complications.

1.  $q$  may itself contain adjustable parameters;  $q'$  contains one more but reduces to  $q$  if and only if the extra parameter is zero. We shall refer to the adjustable parameters present on  $q$  as old parameters, those present on  $q'$  but not on  $q$  as new parameters.

2.  $q'$  may contain more than one new parameter.

3. Two sets of observations may be considered. They are supposed derived from laws of the same form, but it is possible that one or more parameters in the laws have different values. Then  $q$  is the hypothesis

that the parameters have the same value in the two sets,  $q'$  that at least one of them has different values.

4. It may be already established that some parameters have different values on the two laws, but the question may be whether some further parameter differs. For instance, the two sets of data may both be derived from normal laws, and the standard errors may already be known to differ; but the question of the agreement of the true values remains open. This state of affairs is particularly important when a physical constant has been estimated by totally different methods and we want to know whether the results are consistent.

5. More than two sets of observations may have to be compared. Several sets may agree, but one or more may be found to differ from the consistent sets by amounts that would be taken as significant if they stood by themselves. But in such cases we are picking out the largest discrepancy, and a discrepancy of any amount might arise by accident if we had enough sets of data. Some allowance for selection is therefore necessary in such cases.

**5.01. Treatment of old parameters.** Suppose that there is one old parameter  $\alpha$ ; the new parameter is  $\beta$ , and is 0 on  $q$ . In  $q'$  we could replace  $\alpha$  by  $\alpha'$ , any function of  $\alpha$  and  $\beta$ ; but to make it explicit that  $q'$  reduces to  $q$  when  $\beta = 0$  we shall require that  $\alpha' = \alpha$  when  $\beta = 0$ . Suppose that  $\alpha'$  is chosen so that  $\alpha'$  and  $\beta$  are orthogonal parameters in the sense of 4.31; take

$$P(q d\alpha | H) = h(\alpha) d\alpha, \quad P(q' d\alpha' d\beta | H) = h(\alpha') d\alpha' f(\beta, \alpha') d\beta, \quad (1)$$

where 
$$\int f(\beta, \alpha') d\beta = 1. \quad (2)$$

For small changes of  $\alpha'$  and  $\beta$ ,

$$J = g_{\alpha\alpha} d\alpha'^2 + g_{\beta\beta} d\beta^2. \quad (3)$$

If  $n$  is large, we get maximum likelihood estimates  $a$  and  $b$  for  $\alpha'$  and  $\beta$ , and

$$P(dadb | q\alpha H) \propto \frac{1}{2\pi} \exp\left[-\frac{1}{2}n\{g_{\alpha\alpha}(\alpha-a)^2 + g_{\beta\beta}b^2\}\right], \quad (4)$$

$$P(dadb | q'\alpha'\beta H) \propto \frac{1}{2\pi} \exp\left[-\frac{1}{2}n\{g_{\alpha\alpha}(\alpha'-a)^2 + g_{\beta\beta}(\beta-b)^2\}\right]. \quad (5)$$

Hence 
$$P(q | abH) \propto \int h(\alpha) \exp\left[-\frac{1}{2}n\{g_{\alpha\alpha}(\alpha-a)^2 + g_{\beta\beta}b^2\}\right] d\alpha$$

$$\propto h(a) \sqrt{\left(\frac{2\pi}{ng_{\alpha\alpha}}\right)} \exp\left(-\frac{1}{2}ng_{\beta\beta}b^2\right), \quad (6)$$

$$P(q' | abH) \propto \iint h(\alpha') f(\beta, \alpha') \exp[-\frac{1}{2}n\{g_{\alpha\alpha}(\alpha' - a)^2 + g_{\beta\beta}(\beta - b)^2\}] d\alpha' d\beta \\ \propto h(a) f(b, a) \frac{2\pi}{n\sqrt{(g_{\alpha\alpha}g_{\beta\beta})}}, \quad (7)$$

$$K \doteq \frac{1}{f(b, a)} \sqrt{\left(\frac{ng_{\beta\beta}}{2\pi}\right)} \exp(-\frac{1}{2}ng_{\beta\beta}b^2). \quad (8)$$

This is of the same form as 5.0 (10). To the accuracy of this approximation  $h(\alpha)$  is irrelevant. It makes little difference to  $K$  whether we have much or little previous information about the old parameter.  $f(\beta, \alpha')$  is a prior probability density for  $\beta$  given  $\alpha'$ .

If  $\alpha''$  also reduces to  $\alpha$  when  $\beta = 0$ , but is not orthogonal to  $\beta$  for small values of  $\beta$ , we may take

$$\alpha'' = \alpha' + \lambda\beta. \quad (9)$$

If instead of (1) we take

$$P(q' d\alpha'' d\beta | H) = h(\alpha'') f(\beta, \alpha'') d\alpha'' d\beta \quad (10)$$

we are led to

$$P(q' | abH) \propto \iint h(\alpha'') f(\beta, \alpha'') \exp[-\frac{1}{2}n\{g_{\alpha\alpha}(\alpha' - a)^2 + g_{\beta\beta}(\beta - b)^2\}] d\alpha' d\beta \\ \doteq h(a + \lambda b) f(b, a + \lambda b) \frac{2\pi}{n\sqrt{(g_{\alpha\alpha}g_{\beta\beta})}} \quad (11)$$

provided now that  $h$  varies slowly. There will be little change in  $K$  if  $b$  is small and we have little previous information about  $\alpha''$ ; so that the condition that old parameters shall be taken orthogonal to the new ones makes little difference to the results. But if there is much previous information about  $\alpha''$  we may have to take account of the variation of  $h(\alpha'')$  in the range where the exponential is not small, and the disturbance of the result may be considerable.

There is therefore no difficulty in principle in allowing for old parameters. If previous considerations, such as a definite hypothesis or even a consistent model, suggest a particular way of specifying them on  $q'$ , we may use it. If not, we can take them orthogonal to the new one, because this automatically satisfies the condition that the parameter  $\alpha'$  that replaces  $\alpha$  on  $q'$  shall reduce to  $\alpha$  when  $\beta = 0$ ; then the prior probability of  $\alpha$  on  $q$  can be immediately adapted to give a suitable one for  $\alpha'$  on  $q'$ . In these cases the result will be nearly independent of previous information about the old parameters.

In the first edition of this book I made it a rule that old parameters on  $q'$  should be defined in such a way that they would have maximum likelihood estimates independent of the new parameter. This was rather unsatisfactory because in estimation problems maximum likelihood

arises as a derivative principle, as an approximation to the principle of inverse probability. It seemed anomalous that it should appear, apparently as a postulate, in the principles of significance tests. We now see that it is unnecessary, but that the notion of orthogonality leads to a specially convenient statement of the method; and orthogonal parameters satisfy the rule of the first edition to the accuracy required.

**5.02. Required properties of  $f(\alpha)$ .** To arrive at quantitative results we need to specify the function  $f(\alpha)$  of 5.0 or  $f(\beta, \alpha)$  of 5.01. It might appear that on  $q'$  the new parameter is regarded as unknown and therefore that we should use the estimation prior probability for it. But this leads to an immediate difficulty. Suppose that we are considering whether a location parameter  $\alpha$  is 0. The estimation prior probability for it is uniform, and subject to 5.0(3) we should have to take  $f(\alpha) = 0$ , and  $K$  would always be infinite. We must instead say that the mere fact that it has been suggested that  $\alpha$  is zero corresponds to some presumption that it is fairly small. Then we can make a test with any form of  $f(\alpha)$  whose integral converges. But it must not converge too fast, or we shall find that the null hypothesis can never be sufficiently decisively rejected. We shall deal with this explicitly later. At present we need only remark that the effect of a suggestion that  $\alpha = 0$ , if it has to be rejected, implies much less evidence against large values of  $\alpha$  than would be provided by a single observation that would give a maximum likelihood solution  $\alpha = 0$ . In cases where a single observation would not give strong evidence against large values of  $\alpha$ , it will be enough to use the estimation prior probability.

The situation appears to be that when a suggestion arises that calls for a significance test there may be very little previous information or a great deal. In sampling problems the suggestion that the whole class is of one type may arise before any individual at all has been examined. In the establishment of Kepler's laws several alternatives had to be discussed and found to disagree wildly with observation before the right solutions were found, and by the time when perturbations began to be investigated theoretically the extent of departures from Kepler's laws was reasonably well known, and well beyond the standard error of one observation. In experimental physics it usually seems to be expected that there will be systematic error comparable with the standard error of one observation. In much modern astronomical work effects are deliberately sought when previous information has shown that they may be of the order of a tenth of the standard error of one observation, and consequently there is no hope of getting a decision one way or the

other until some hundreds of observations have been taken. In any of these cases it would be perfectly possible to give a form of  $f(\alpha)$  that would express the previous information satisfactorily, and consideration of the general argument of 5.0 will show that it would lead to common-sense results, but they would differ in scale. As we are aiming chiefly at a theory that can be used in the early stages of a subject, we shall not at present consider the last type of case; we shall see that the first two are covered by taking  $f(\alpha)$  to be of the form  $C/(1+\alpha^2/\sigma^2)$ .

**5.03. Comparison of two sets of observations.** Let two sets of observations, of numbers  $n_1, n_2$ , be derived from laws that agree in parameters  $\alpha_1, \dots, \alpha_m$ , but possibly differ in a parameter  $\alpha_{m+1}$ . Let the values of  $\alpha_{m+1}$  in the two be  $\beta_1, \beta_2$ . The standard error of  $\beta_1 - \beta_2$  as found in an estimation problem would be

$$s = O\left(\frac{n_1 + n_2}{n_1 n_2}\right)^{1/2}. \quad (1)$$

Then the first factor in 5.0 (10) will be

$$O\left(\frac{n_1 n_2}{n_1 + n_2}\right)^{1/2} \frac{1}{f(0)}. \quad (2)$$

Now if  $n_2$  is very large compared with  $n_1$  we are practically comparing the estimate of  $\beta_1$  with an accurately determined value, and (2) should be  $O(n_1^{1/2})$ . It is, provided  $f(0)$  is independent of  $n_1$ , and by symmetry of  $n_2$ .

This principle is not satisfied by two of the tests given in the first edition of this book: comparison of two series of measures when the standard errors are equal (5.51); and comparison of two standard errors (5.53). In these the factor in question was  $O(n_1 + n_2)^{1/2}$ . The prior probability of the difference of the parameters on the alternative hypothesis in these can be seen on examination to depend on  $n_1/n_2$ . The method was based on somewhat artificial partitions of expectations.

**5.04. Selection of alternative hypotheses.** So far we have considered the comparison of the null hypothesis with a simple alternative, which could be considered as likely as the null hypothesis. Sometimes, however, the use of  $\chi^2$  or  $z$ , or some previous consideration, suggests that some one of a group of alternative hypotheses may be right without giving any clear indication of which. For instance, the chief periods in the tides and the motion of the moon were detected by first noticing that the observed quantity varied systematically and then examining the departures in detail. In such a case (we are supposing for a moment that we are in a pre-Newtonian position without a gravitational theory

to guide us) the presence of one period by itself would give little or no reason to expect another. We may say that the presence of various possible periods gives alternative hypotheses  $q_1, q_2, \dots$ , whose disjunction is  $q'$ . They are mutually irrelevant, and therefore not exclusive. Suppose then that the alternatives are  $m$  in number, all with probability  $k$  initially, and that

$$P(q | H) = P(q' | H) = \frac{1}{2}. \quad (1)$$

Since we are taking the various alternatives as irrelevant the probability that they are all false is  $(1-k)^m$ . But the proposition that they are all false is  $q$ ; hence

$$(1-k)^m = \frac{1}{2}, \quad (2)$$

$$k = 1 - 2^{-1/m} \doteq \frac{1}{m} \log 2, \quad (3)$$

if  $m$  is large. Thus, if we test the hypothesis  $q_1$  separately we shall have

$$\frac{P(q | H)}{P(q_1 | H)} = \frac{1}{2k} \doteq \frac{m}{2 \log 2} = 0.7m, \quad (4)$$

nearly. If  $K$  is found by taking  $P(q | H) = P(q_1 | H)$ , we can correct for selection by multiplying  $K$  by  $0.7m$ .

Where the data are frequencies or the values of a continuous quantity at a set of discrete values of the argument, a finite number of Fourier amplitudes suffice to express the whole of the data exactly, and the procedure would be to test these in order, preferably beginning with the largest. An intermediate real period would contribute to more than one estimated amplitude, and the true period could then be estimated by comparison of adjacent amplitudes.†

Where the dependent variable is a continuous function and we have a continuous record of it, neighbouring values are correlated in any circumstances. It would be wrong to treat neighbouring values as subject to independent errors. The null hypothesis would be more like a statement that a finite number of values are assigned at random and that the intermediate ones are represented by the interpolation function. The problem is a case of what is now known as serial correlation. A method that is often used is to divide the interval into several, do separate analyses for each, and estimate an uncertainty by comparison.

In practice it is rather unusual for a set of parameters to arise in such a way that each can be treated as irrelevant to the presence of

† This method differs appreciably from the 'periodogram' method of Schuster, which may miss some periods altogether and estimate amplitudes of others that lie too close together to be independent. It is essentially due to H. H. Turner. For details see H. and B. S. Jeffreys, *Methods of Mathematical Physics*, pp. 400, 421.



any other. Even in the above case each period means two new parameters, representing the coefficients of a sine and cosine; the presence of a period also would usually suggest the presence of its higher harmonics. More usual cases are where one new parameter gives inductive reason, but not demonstrative reason, for expecting another, and where some parameters are so closely associated that one could hardly occur without the others.

The former case is common in the discussion of estimates of a physical constant from different sets of data, to see whether there are any systematic differences between them. The absence of such differences can be taken as the null hypothesis. But if one set is subject to systematic error, that gives some reason to expect that others are too. The problem of estimating the numbers of normal and abnormal sets is essentially one of sampling, with half the prior probability concentrated at one extreme; but we also want to say, as far as possible, which are the abnormal sets. The problem is therefore to draw the line, and since  $K$  depends chiefly on  $\chi^2$  it is convenient to test the sets in turn in order of decreasing contributions to  $\chi^2$ . If at any stage we are testing the  $p$ th largest contribution ( $p > 1$ ),  $p-1$  have already been found abnormal. Suppose that  $s$  have been found normal. Then at this stage both extreme possibilities have been excluded and the ratio of the prior probabilities that the  $p$ th largest contribution is normal or abnormal is  $(s+1)/p$ , by Laplace's theory. In practice, if there are  $m$  sets,  $s$  can be replaced by  $m-p$ ; for if the  $p$ th is the smallest abnormal contribution,  $s$  will be equal to  $m-p$ , so that the line will be drawn in the right place. Hence  $K$  as found in a simple test must be multiplied by  $(m-p+1)/p$ . We can then begin by testing the extreme departure, taking  $p = 1$ ,  $s = m-1$ , and therefore multiplying  $K$  by  $m$ . If the corrected  $K$  is less than 1 we can proceed to the second, multiplying this time by  $(m-1)/2$ , and so on. There is a complication, however, if the first passes the test and the second does not. For the multiplication by  $m$  supposes both extreme cases excluded already. In testing the first we have not yet excluded  $q$ , and if we find no other abnormal cases the question will arise whether we have not after all decided wrongly that the first was abnormal. This can be treated as follows. The factor  $m$  arises from Laplace's theory, which makes the prior probabilities of  $q$  (no abnormal cases) and  $q'$  (at least one abnormal case) in the ratio 1 to  $m$ . At the outset, however, we are taking these probabilities equal, and therefore we should multiply  $K$  by  $m^2$  instead of  $m$ . We can start with  $m$ ; but if the second departure tested does not give a corrected  $K$  less than 1

we should return to the first and apply a factor  $m^2$  instead of  $m$ . It is best to proceed in this order, because to apply the factor  $m^2$  at the first step might result in the acceptance of  $q$  at once and prevent any use from being made of the second largest contribution to  $\chi^2$ , which might be nearly as large as the first.

In comparison with the case where the suggested abnormalities are irrelevant, the correcting factors to  $K$  here are somewhat larger for testing the largest contributions to  $\chi^2$ , and smaller for the smaller ones.

The need for such allowances for selection of alternative hypotheses is serious. If a single hypothesis is set up for test, the critical value may be such that there would be a probability of 0.05 that it would be exceeded by accident even if  $q$  was true. We have to take such a risk if we are to have any way of detecting a new parameter when it is needed. But if we tested twenty new parameters according to the same rule the probability that the estimate of one would exceed the critical value by accident would be 0.63. In twenty trials we should therefore expect to find an estimate giving  $K < 1$  even if the null hypothesis was correct, and the finding of 1 in 20 is no evidence against it. If we persist in looking for evidence against  $q$  we shall always find it unless we allow for selection. The first quantitative rule for applying this principle was due, I think, to Sir G. T. Walker;† analogous recommendations are made by Fisher.‡

**5.1. Test of whether a suggested value of a chance is correct.** An answer in finite terms can be obtained in the case where the parameter in question is a chance, and we wish to know whether the data support or contradict a value suggested for it. Suppose that the suggested value is  $p$ , that the value on  $q'$ , which is so far unknown, is  $p'$ , and that our data consist of a sample of  $x$  members of one type and  $y$  of the other. Then on  $q'$ ,  $p'$  may have any value from 0 to 1. Thus

$$P(q | H) = \frac{1}{2}, \quad P(q' | H) = \frac{1}{2}, \quad P(dp' | q', H) = dp', \quad (1)$$

$$\text{whence} \quad P(q', dp' | H) = \frac{1}{2} dp'. \quad (2)$$

Also, if  $\theta$  denotes the observational evidence,

$$P(\theta | qH) = {}^{x+\nu}C_x p^x (1-p)^\nu, \quad (3)$$

$$P(\theta | q', p', H) = {}^{x+\nu}C_x p'^x (1-p')^\nu; \quad (4)$$

whence

$$P(q | \theta H) \propto p^x (1-p)^\nu, \quad (5)$$

$$P(q', dp' | \theta H) \propto p'^x (1-p')^\nu dp', \quad (6)$$

† *Q. J. R. Met. Soc.* 51, 1925, 337-48.

‡ *Statistical Methods for Research Workers*, 1936, pp. 65-6.

and by integration

$$P(q' | \theta H) \propto \int_0^1 p'^x (1-p')^y dp' = \frac{x!y!}{(x+y+1)!}. \quad (7)$$

$$\text{Hence} \quad K = \frac{P(q | \theta H)}{P(q' | \theta H)} = \frac{(x+y+1)!}{x!y!} p^x (1-p)^y. \quad (8)$$

If  $x$  and  $y$  are large, an approximation by Stirling's theorem gives

$$K \div \left\{ \frac{x+y}{2\pi p(1-p)} \right\}^{1/2} \exp \left\{ -\frac{\{x-p(x+y)\}^2}{2(x+y)p(1-p)} \right\}. \quad (9)$$

The following table indicates how  $K$  varies with  $x$  and  $y$  when these are small and  $p = \frac{1}{2}$ ; that is, if we are testing whether a chance is even:

$x$	$y$	$K$	$x$	$y$	$K$	$x$	$y$	$K$
1	0	1	1	1	$\frac{3}{2}$	2	2	$\frac{15}{8}$
2	0	$\frac{3}{4}$	2	1	$\frac{3}{2}$	3	3	$\frac{35}{16}$
3	0	$\frac{1}{2}$	3	1	$\frac{5}{4}$	4	4	$\frac{315}{128}$
4	0	$\frac{1}{16}$	4	1	$\frac{15}{8}$	5	5	$\frac{3283}{256}$
5	0	$\frac{3}{16}$	5	1	$\frac{31}{32}$			

None of these ratios is very decisive, and a few additional observations can make an appreciable change. The most decisive is for  $x = 5$ ,  $y = 0$ , and even for that the odds in favour of a bias are only those in favour of picking a white ball at random out of a box containing sixteen white ones and three black ones—odds that would interest a gambler, but would be hardly worth more than a passing mention in a scientific paper. We cannot get decisive results one way or the other from a small sample.

The result  $K = 1$  for  $x = 1$ ,  $y = 0$  is interesting. The first member sampled is bound to be of one type or the other, whether the chance is  $\frac{1}{2}$  or not, and therefore we should expect it to give no information about the existence of bias. This is checked by the result  $K = 1$  for this case. Similarly, if  $x = y$ , we have

$$K = \frac{(2x+1)!}{x!x!} \left(\frac{1}{2}\right)^{2x},$$

and if  $y$  is increased to  $x+1$

$$K = \frac{(2x+2)!}{x!(x+1)!} \left(\frac{1}{2}\right)^{2x+1},$$

which is the same. Thus if at a certain stage the sample is half and half, the next member, which is bound to be of one type or the other, gives no new information.

This holds only if  $p = \frac{1}{2}$ . If  $p = \frac{3}{4}$ ,  $x = 1$ ,  $y = 0$ , we get  $K = \frac{3}{2}$ ; but if  $x = 0$ ,  $y = 1$  we get  $K = \frac{1}{2}$ . This is because, if the less likely

event on  $q$  comes off at the first trial, it is some evidence against  $q$ , and if the likely one comes off it is evidence for  $q$ . This is reasonable.

For  $p = \frac{1}{2}$ ,  $K$  first becomes  $< 0.1$  for  $x = 7, y = 0$ , and first becomes  $> 10$  for  $x = y = 80$ . To get this amount of support for an even chance requires as much evidence as would fix a ratio found by sampling within a standard error of  $(\frac{1}{2} \cdot \frac{1}{2}/160)^{1/2} = 0.04$ . It is therefore possible to obtain strong evidence against  $q$  with far fewer observations than would be needed to give equally strong evidence for it. This is a general result and corresponds to the fact that while the first factor in 5.1 (9) increases only like  $n^{1/2}$ , the second factor, for a given value of  $p'$ , will decrease like  $\exp[-\alpha n(p' - p)^2]$ , where  $\alpha$  is a moderate constant. We notice too that the expectations of  $x$  and  $y$  on  $q$  are  $(x+y)p$  and  $(x+y)(1-p)$ ; so that

$$\chi^2 = \frac{\{x - (x+y)p\}^2}{(x+y)p} + \frac{\{y - (x+y)(1-p)\}^2}{(x+y)(1-p)} = \frac{\{x - (x+y)p\}^2}{(x+y)p(1-p)} \quad (10)$$

and the exponential factor is  $\exp(-\frac{1}{2}\chi^2)$ . This is a general result for problems where the standard error is fixed merely by the numbers of observations.

A remarkable series of experiments was carried out by W. F. R. Weldon† to test the bias of dice. The question here was whether the chance of a 5 or a 6 was genuinely  $\frac{1}{3}$ . In 315672 throws, 106602 gave a 5 or a 6. The ratio is 0.337699, suggesting an excess chance of 0.004366. We find

$$K = \left(\frac{315672}{2\pi \cdot \frac{1}{3} \cdot \frac{2}{3}}\right)^{1/2} \exp\left[-\frac{1}{2} \frac{315672 \times 0.004366^2}{\frac{1}{3} \cdot \frac{2}{3}}\right] \\ = 476 \exp[-13.539] = 6.27 \times 10^{-4},$$

so that the odds are about 1600 to 1 in favour of a small bias. Extreme care was taken that a possible bias in the conditions of throwing should be eliminated; the dice were actually rolled, twelve at a time, down a slope of corrugated cardboard. The explanation appears to be that in the manufacture of the dice small pits are made in the faces to accommodate the marking material, and this lightens the faces with 5 or 6 spots, displacing the centre of gravity towards the opposite sides and increasing the chance that these faces will settle upwards.

The formula for testing an even chance is of great use in cases where observations are given in a definite order, and there is a question whether they are independent. If we have a set of residuals against an assigned formula, and they represent only random variation, each is independent of the preceding ones, and the chances of a persistence and

† Quoted by Pearson, *Phil. Mag.* 50, 1900.

a change of sign are equal. We can therefore count the persistences and changes, and compare the numbers with an even chance. If a number of functions have been determined from the data, each introduces one change of sign, so that the number of changes should be reduced by the number of parameters determined. Similarly, if we have a series of events of two types and they are independent, the same rule will hold. We may try it on the set of possible results of random sampling given in 2.13. For the series obtained by coin-tossing we have 7 persistences and 13 changes, giving nearly

$$K = \left( \frac{2 \times 20}{\pi} \right)^{1/2} \exp(-0.9) = 1.5.$$

This may be accepted as a random series. The second series also gives 7 persistences and 13 changes and the same value of  $K$ ; but if we compared each observation with one three places before it we should have 18 persistences with no change at all. The next two each give 20 persistences and  $K = 2 \times 10^{-5}$ . The last gives 20 persistences and 5 changes, and  $K = \frac{1}{23}$  nearly. Thus even with these rather short series the simple test by counting persistences and changes gives the right result immediately in four cases out of five, and in the other it would give it after attention to special types of non-random arrangement, possibly with allowance for selection. The test, however, does not necessarily make use of the whole of the information in the data. It is a convenient and rapid way of detecting large departures, but often fails for small ones that would be revealed by a method that goes more into detail.

**5.11. Simple contingency.** Suppose that a large population is sampled with respect to two properties  $\phi$  and  $\psi$ . There are four alternative combinations of properties. The probability of a member having any pair may be a chance, or the population may be large enough for it to be considered as one. Then the alternatives, the sampling numbers, and the chances may be shown as follows:

$$\begin{pmatrix} \phi.\psi & \phi.\sim\psi \\ \sim\phi.\psi & \sim\phi.\sim\psi \end{pmatrix}, \quad \begin{pmatrix} x & y \\ x' & y' \end{pmatrix}, \quad \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix}.$$

The question is, are  $\phi$  and  $\psi$  associated? that is, are the chances out of proportion? If they are in proportion we have hypothesis  $q$ , that

$$p_{11}p_{22} = p_{12}p_{21}. \quad (1)$$

Whether they are in proportion or not, we can consider the chance of

a member having the property  $\phi$ ; let this be  $\alpha$ , and the chance of  $\psi$ ,  $\beta$ .

Putting  $1 - \alpha = \alpha'$ ,  $1 - \beta = \beta'$ , (2)

we have on  $q$  
$$\begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix} = \begin{pmatrix} \alpha\beta & \alpha\beta' \\ \alpha'\beta & \alpha'\beta' \end{pmatrix}. \quad (3)$$

On  $q'$ , since  $\alpha$  and  $\beta$  are already defined and their amounts have nothing to do with whether  $\phi$  and  $\psi$  are associated, the chances can differ only in such a way that the row and column totals are unaltered; hence there is a number  $\gamma$  such that the set of chances is

$$\begin{pmatrix} \alpha\beta + \gamma & \alpha\beta' - \gamma \\ \alpha'\beta - \gamma & \alpha'\beta' + \gamma \end{pmatrix}, \quad (4)$$

and this is general, since any set of chances, subject to their sum being 1, can be represented by a suitable choice of  $\alpha$ ,  $\beta$ ,  $\gamma$ . Since  $\alpha$  and  $\beta$  are chances, zero and unit values excluded, we have

$$P(q d\alpha d\beta | H) = P(q' d\alpha d\beta | H) = \frac{1}{2} d\alpha d\beta. \quad (5)$$

Also

$$p_{11} p_{22} - p_{12} p_{21} = \gamma, \quad (6)$$

$$\frac{\partial(p_{11}, p_{12}, p_{21})}{\partial(\alpha, \beta, \gamma)} = \begin{vmatrix} \beta & \beta' & -\beta \\ \alpha & -\alpha & \alpha' \\ 1 & -1 & -1 \end{vmatrix} = 1. \quad (7)$$

Since  $\gamma$  is linearly related to the separate chances it is natural to take its prior probability as uniformly distributed. But  $\alpha$  and  $\beta$  impose limits on the possible values of  $\gamma$ . With a mere rearrangement of the table we can make  $\alpha < \alpha'$ ,  $\beta < \beta'$ ,  $\alpha\beta' < \alpha'\beta$ . Since

$$\alpha'\beta - \alpha\beta' = \beta - \alpha, \quad (8)$$

this makes  $\alpha$  the smallest of  $\alpha$ ,  $\beta$ ,  $\alpha'$ ,  $\beta'$ . Then the possible values of  $\gamma$  lie between  $-\alpha\beta$  and  $\alpha\beta'$ , since no chance can be negative; and

$$P(d\gamma | q', \alpha, \beta, H) = d\gamma / \alpha. \quad (9)$$

Hence

$$P(q' d\alpha d\beta d\gamma | H) = \frac{1}{2} d\alpha d\beta d\gamma / \alpha. \quad (10)$$

In ranges where  $\alpha$  is not the smallest of  $\alpha$ ,  $\beta$ ,  $\alpha'$ ,  $\beta'$ , it must be replaced in the denominator by the smallest.

Now the chances of getting the observed results in their actual order are in each case  $p_{11}^x p_{12}^y p_{21}^{x'} p_{22}^{y'}$ . Hence

$$P(q d\alpha d\beta | \theta H) \propto \alpha^{x+\nu} \alpha'^{x'+\nu'} \beta^{x+x'} \beta'^{y+y'} d\alpha d\beta, \quad (11)$$

$$P(q' d\alpha d\beta d\gamma | \theta H) \propto (\alpha\beta + \gamma)^x (\alpha\beta' - \gamma)^y (\alpha'\beta - \gamma)^{x'} (\alpha'\beta' + \gamma)^{y'} d\alpha d\beta d\gamma / \alpha. \quad (12)$$

Integrating the former we get

$$P(q | \theta H) \propto \frac{(x+y)! (x'+y')! (x+x')! (y+y')!}{\{(x+x'+y+y'+1)!\}^2}. \quad (13)$$

We have 
$$\frac{\partial(\alpha, p_{11}, p_{21})}{\partial(\alpha, \beta, \gamma)} = -1, \quad (14)$$

and the integral of (12) is nearly

$$\begin{aligned} P(q' | \theta H) &\propto \int_0^1 \int_0^\alpha \int_0^{1-\alpha} p_{11}^\alpha (\alpha - p_{11})^\nu p_{21}^{\alpha'} (1 - \alpha - p_{21})^{\nu'} d\alpha dp_{11} dp_{21} / \alpha \\ &= \int_0^1 \frac{x! y!}{(x+y+1)!} \alpha^{x+\nu} \frac{x'! y'!}{(x'+y'+1)!} (1-\alpha)^{x'+\nu'+1} d\alpha \\ &= \frac{x! y! x'! y'!}{(x+y+1)(x+y+x'+y'+2)!}, \end{aligned} \quad (15)$$

$$K = \frac{(x+y+1)! (x'+y')! (x+x')! (y+y')!}{x! y! x'! y'! (x+y+x'+y'+1)!} (x+y+x'+y'+2). \quad (16)$$

An approximation has been made in allowing  $\alpha$  to range from 0 to 1, since  $\alpha < \beta < \frac{1}{2}$ ; but if  $x+y$  is the smallest total,  $\alpha$  is about

$$\frac{x+y \pm \sqrt{(x+y)}}{x'+y'+x+y},$$

and the contribution from the extra range is exponentially small unless  $\alpha$  and  $\beta$  are nearly equal. The exact procedure would be to replace  $\alpha$  by  $\beta$ ,  $\alpha'$ , or  $\beta'$  in ranges where  $\alpha$  is not the smallest; thus we have very slightly underestimated  $P(q' | \theta H)$  and overestimated  $K$ .

If  $x'$  and  $y'$  are very large compared with  $x$  and  $y$ , the chance of  $\psi$ , given  $\sim \phi$ , is very accurately given by  $x'/(x'+y')$ . Replacing this by  $p$  we have

$$K = \frac{(x+y+1)!}{x! y!} \frac{x^x y^y \nu^\nu}{(x'+y')^{x+\nu}} = \frac{(x+y+1)!}{x! y!} p^x (1-p)^\nu, \quad (17)$$

which is the same as 5.1 (8). This was to be expected. The present form is a little more accurate than a previous one,<sup>†</sup> in which I integrated without reference to the variation of  $p_{11} + p_{12}$ , replacing the latter after integration by its most probable value. The result was that the extra factor  $x+y+1$  was replaced by  $x+y$ . The difference is trivial, but will give an idea of the amount of error introduced by the procedure of integrating the factors with large indices and replacing those with small indices by their most probable values at the end of the work.

If  $x$ ,  $y$ ,  $x'$ ,  $y'$  are all large we can approximate by Stirling's formula; then

$$K = \left\{ \frac{(x+y+x'+y')^3 (x+y)}{2\pi (x+x') (x'+y') (y+y')} \right\}^{1/2} \exp \left[ -\frac{1}{2} \frac{(x+y+x'+y') (xy' - x'y)^2}{(x+y)(x+x')(x'+y')(y+y')} \right], \quad (18)$$

<sup>†</sup> *Proc. Roy. Soc. A*, 162, 1937, 479-95.

where  $x+y$  is defined to be the smallest of the four row and column sums. The exponential factor is  $\exp(-\frac{1}{2}\chi^2)$ . For if we put

$$N = x+y+x'+y',$$

the four expectations on  $q$ , given the row and column totals, are

$$\frac{(x+y)(x+x')}{N}, \quad \frac{(x+y)(y+y')}{N}, \quad \frac{(x+x')(x'+y')}{N}, \quad \frac{(x'+y')(y+y')}{N},$$

and 
$$x - \frac{(x+y)(x+x')}{N} = \frac{xy' - x'y}{N},$$

the other residuals being equal or equal and opposite to this. Hence

$$\begin{aligned} \chi^2 &= \left( \frac{xy' - x'y}{N} \right)^2 \times \\ &\quad \times \left\{ \frac{N}{(x+y)(x+x')} + \frac{N}{(x+y)(y+y')} + \frac{N}{(x+x')(x'+y')} + \frac{N}{(x'+y')(y+y')} \right\} \\ &= \frac{N(xy' - x'y)^2}{(x+y)(x+x')(y+y')(x'+y')}. \end{aligned} \quad (19)$$

**5.12. Comparison of samples.** In the last problem the only restriction on the sample is its total number  $N$ . If  $\phi$  is a rare property, we may require a prohibitively large sample to make  $x$  and  $y$  large enough to give a decisive test one way or the other. But it may be possible to arrange the sampling so that  $x+y$  and  $x'+y'$  are both large enough to be useful, without violating the condition that, given either  $\phi$  or  $\sim\phi$ , a member has the same chance of being included in the sample whether it has  $\psi$  or  $\sim\psi$ . Thus, if we want to know whether red hair is more frequent among Englishmen or Scotsmen, we might take a sample at random from the population of London, and classify the results in a  $2 \times 2$  contingency table. But if such a sample is to contain enough Scotsmen to give much information it will contain more Englishmen than it is practicable to classify. We can, however, proceed in two other ways. We can sample at random till we have, say, 200 Englishmen, and after that we can ignore further Englishmen and count Scotsmen only, until we have a suitable number of the latter. Or we can take a random sample of 200 Englishmen from London, and another of 200 Scotsmen from Perth, and compare the two samples. If  $\phi$  is the property 'Scottish' and  $\sim\phi$  'English', these methods do not attempt to provide information about  $\alpha$ , but replace it by two sample totals  $x+y$  and  $x'+y'$  determined for convenience.



On hypothesis  $q$  the chance of  $\psi$  is the same, given either  $\phi.H$  or  $\sim\phi.H$ . Call this  $\beta$ . Then

$$P(d\beta | qH) = d\beta, \quad (1)$$

$$P(\theta | q, \beta, H) = \beta^{x+x'}(1-\beta)^{y+y'}, \quad (2)$$

and  $(x+y+x'+y')\beta$  is the expectation of  $\psi$ 's in a sample of  $x+y+x'+y'$  in all. To have a valid standard of comparison, if  $p$  and  $p'$  are the chances of  $\psi$  on  $q'.\phi H$  and on  $q'.\sim\phi.H$ , we must define a  $\beta$  by

$$N\beta = (x+y+x'+y')\beta = (x+y)p + (x'+y')p', \quad (3)$$

so that the left side will still be the expectation of the number of  $\psi$ 's in the two samples together.  $\beta$  has the property that it is orthogonal to  $p-p'$ . Then both  $p$  and  $p'$  must be between 0 and 1. Within the permitted range for  $p, p'$  for a given  $\beta$  can have values from  $N\beta/(x'+y')$  to  $(N\beta-x-y)/(x'+y')$ . But the most probable value of  $N\beta$  will be nearly  $x+x'$ . The former value will then be permissible if  $x < y'$ , and the latter if  $x' > y$ , and if these are satisfied there will be no further restriction. Then

$$P(d\beta | q', H) = d\beta, \quad (4)$$

$$P(dp | q', \beta, H) = dp, \quad (5)$$

$$\frac{\partial(\beta, p)}{\partial(p', p)} = \frac{x'+y'}{N}, \quad (6)$$

$$P(\theta | p, p', q', H) = p^x(1-p)^y p'^{x'}(1-p')^{y'}. \quad (7)$$

Hence

$$P(q | \theta H) \propto \int_0^1 \beta^{x+x'}(1-\beta)^{y+y'} d\beta = \frac{(x+x')!(y+y')!}{(x+x'+y+y'+1)!}, \quad (8)$$

$$\begin{aligned} P(q' | \theta H) &\propto \int_0^1 \int_0^1 p^x(1-p)^y p'^{x'}(1-p')^{y'} d\beta dp \\ &= \frac{x'+y'}{N} \iint p^x(1-p)^y p'^{x'}(1-p')^{y'} dp dp' \\ &\doteq \frac{x'+y'}{N} \frac{x!y!}{(x+y+1)!} \frac{x'!y'!}{(x'+y'+1)!}, \end{aligned} \quad (9)$$

$$K = \frac{(x+y+1)!}{x!y!} \frac{(x'+y')!(x+x')!(y+y')!(x'+y'+1)N}{x'!y'!(N+1)!(x'+y')}, \quad (10)$$

which differs from 5.11 (16) only by quantities of order  $1/(x'+y')$ .

**5.13.** If  $x' < y$  (more strictly, if  $(x'+y')p' < (x+y)(1-p)$ ) the pos-

sible values of  $p$  impose a further restriction, since the largest possible value of  $p$  is now  $N\beta/(x+y)$ . Then (4) and (6) still hold, but

$$P(dp | \beta, q', H) = \frac{x+y}{N\beta} dp, \quad (1)$$

and we are led to

$$P(q' | \theta H) \propto \frac{x+y}{N} \frac{x'+y'}{N} \int \int \frac{1}{\beta} p^x (1-p)^y p'^x (1-p')^y dp dp' \quad (2)$$

and at the maximum of the integrand  $\beta = (x+x')/N$  nearly. Hence

$$K = \frac{(x+x')(x+y)! (x+x')! (y+y')! (x+x')!}{x! y! x'! y'! (x+y+x'+y')!} \quad (3)$$

nearly; and with errors of order  $1/(x+x')$  this is the same as we get by interchanging  $x+x'$  with  $x+y$  in 5.11 (16) and 5.12 (10) according to the altered sign of their difference.

5.14. The actual agreement is rather closer, as we can see by studying the case where  $\beta$  is very small. In this case we may be led to the Poisson rule, and to the rule  $P(d\beta | H) \propto d\beta/\beta$  instead of the usual uniform one. But the discrepancy in the results, such as it is, consists of a replacement of  $x+x'+1$  by  $x+x'$ , and this, if genuine and not merely an error of approximation, should persist when  $y$  and  $y'$  are very large. The range permitted to  $p$  will still be restricted to  $\beta$ , but it is best to insert a function  $f(\beta)$  to generalize the prior probability of  $\beta$ . Then we shall have

$$P(q | \theta H) \propto \int_0^1 f(\beta) \beta^{x+x'} (1-\beta)^{y+y'} d\beta, \quad (1)$$

$$P(q' | \theta H) \propto \int \int \frac{x+y}{N\beta} f(\beta) p^x (1-p)^y p'^x (1-p')^y d\beta dp. \quad (2)$$

If  $y$  and  $y'$  are large and  $p$  and  $p'$  small, these reduce to

$$P(q | \theta H) \propto \int_0^1 f(\beta) \beta^{x+x'} \exp\{-\beta(y+y')\} d\beta, \quad (3)$$

$$P(q' | \theta H) \propto \int \int \frac{y}{y+y'} \frac{f(\beta)}{\beta} p^x p'^x \exp\{-py - p'y'\} d\beta dp, \quad (4)$$

and we have

$$(y+y')\beta = py + p'y', \quad (5)$$

so that we can put

$$yp = (y+y')\beta\eta, \quad y'p' = (y+y')\beta(1-\eta), \quad (6)$$

where the permitted range of  $\eta$  is from 0 to 1. Then

$$P(q' | \theta H) \propto \int_0^1 \int_0^1 f(\beta) \beta^{x+x'} \exp\{-\beta(y+y')\} \eta^x (1-\eta)^{x'} \frac{(y+y')^{x+x'}}{y^x y'^{x'}} d\beta d\eta \quad (7)$$

and the integrals involving  $\beta$  in (3) and (7) are identical whatever the form of  $f(\beta)$ . Hence  $\beta$  gives only an irrelevant factor, and

$$\frac{1}{K} = \frac{(y+y')^{x+x'}}{y^x y'^{x'}} \int_0^1 \eta^x (1-\eta)^{x'} d\eta, \quad (8)$$

$$K = \frac{(x+x'+1)!}{x! x'!} \frac{y^x y'^{x'}}{(y+y')^{x+x'}}, \quad (9)$$

which is correct to  $O(y^{-1}, y'^{-1})$  and is valid subject to the conditions that the Poisson law may be substituted for the binomial. Also it is identical in form with 5.1 (8); thus the agreement of two small estimated chances  $x/(x+y)$  and  $x'/(x'+y')$  can be tested by the same formula as the agreement of a chance  $x/(x+x')$  with a predicted one  $y/(y+y')$ . Thus the difference noted in 5.12 is only an error of approximation. It follows that however the sample may be taken, the proportionality of the chances can be tested by

$$K = \frac{(x+y+1)!}{x! y!} \frac{(x+x')!}{x'! y'!} \frac{(y+y')!}{(x+x'+y+y')!} \frac{(x'+y')!}{(x'+y+y')!} \quad (10)$$

$$\div \left\{ \frac{N^3(x+y)}{2\pi(x+x')(y+y')(x'+y')} \right\}^{1/2} \exp(-\frac{1}{2}\chi^2), \quad (11)$$

where  $x+y$  means the smallest of the four totals; and the error is always of order  $K/(x'+y')$ .

Fisher† quotes from Lange the following data on the convictions of twin brothers or sisters (of like sex) of convicted criminals, according as the twins were monozygotic (identical) or dizygotic (no more alike physically than ordinary brothers or sisters). The contingency table, arranged to satisfy the necessary inequalities, is as follows:

			<i>Monozygotic</i>	<i>Dizygotic</i>
Convicted	.	.	10	2
Not convicted	.	.	3	15

Then

$$K = \frac{13!}{10! 2!} \frac{13! 17! 18!}{3! 15! 30!} = \frac{1}{171},$$

while the less accurate exponential approximation gives  $\frac{1}{189}$ . Thus the

† *Statistical Methods for Research Workers*, 1936, p. 99.

latter, even though Stirling's formula and logarithmic approximation have been applied down to  $2!$  and  $3!$ , is still quite reasonably accurate. What we can infer is that, starting without information about whether there is any difference in criminality between similar and dissimilar twins of criminals, we can assert on the data that the odds on the existence of a difference are about 170 to 1.

Yule and Kendall† quote the following official data on the results of inoculation of cattle with the Spahlinger anti-tuberculosis vaccine. The cattle were deliberately infected with tubercle germs, a set of them having first been inoculated. The table, rearranged, is:

		<i>Died or seriously affected</i>	<i>Not seriously affected</i>
Not inoculated	. . .	8	3
Inoculated	. . .	6	13

$$K = \frac{12!}{8!3!} \frac{14!16!19!}{6!13!30!} = 0.37,$$

the exponential approximation 5.11(18) giving 0.31. The odds are about 3 to 1 that inoculation has a preventive effect.

Tables of factorials are given in Comrie's edition of Barlow's tables; of their logarithms, up to  $n = 100$ , in Milne-Thomson and Comrie, *Standard Four-figure Tables*, Table VI.

The following comparison was undertaken to see whether there is any relation between grammatical gender and psychoanalytic symbolism. The list of symbols in Freud's *Introductory Lectures* was taken as a standard, and the corresponding words were taken from Latin, German, and Welsh dictionaries. All synonyms were included; I considered consulting experts in the languages for the usual words, and using the German words from the original edition of the book, but this, I thought, might introduce a bias, and I preferred in the first place to use the whole of the synonyms. The counts were as follows:

	<i>Latin</i>			<i>German</i>			<i>Welsh</i>	
	<i>M.</i>	<i>F.</i>	<i>N.</i>	<i>M.</i>	<i>F.</i>	<i>N.</i>	<i>M.</i>	<i>F.</i>
Male . . .	27	17	4	31	14	7	45	30
Female . . .	10	37	16	15	29	16	28	29

In the first place we ignore neuters and reduce the matter to three  $2 \times 2$  tables. The respective values of  $\chi^2$  are 15.07, 10.78, and 1.55. Using the approximate formula 5.11(18) we get  $K = 1/296$ ,  $1/30$ , and 3.7 for Latin, German, and Welsh respectively. The phenomenon is so striking

† *Introduction to the Theory of Statistics*, 1938, p. 48.

in the two former that a relation between symbolism and gender in them must be considered established, though we see that it is far from being a complete association. It would be more striking still if we combined all three languages, but many words have been adopted from one to another or from common sources, keeping their genders, and the data would not be independent. The association is somewhat stronger in Latin than in German; this is some evidence against the possibility that Freud was guided by the gender in German in his classification.

The non-significant association in Welsh is comprehensible in relation to the other two languages when we inspect the neuters, for Welsh is a two-gender language like French and the primitive neuters have been made masculine. But we notice both in Latin and German a marked tendency for male symbols to avoid the neuter gender; there is a decided preference to make them feminine rather than neuter. On the other hand, a female symbol is somewhat more likely to be neuter than masculine. But when the neuters are made masculine this effect partly counteracts the association between symbolism and masculine or feminine gender. Thus the failure to detect the association in Welsh is not due to the absence of association but to the fact that the greater parts of two genuine effects have been made to cancel by an etymological rule.

The German rule that diminutives are neuter may provide part of the explanation; the three genders may stand originally for father, mother, and child. But this cannot be pursued further here. The immediate result is that the gender of names of inanimate things is not wholly haphazard.

**5.15. Test for consistency of two Poisson parameters.** It may happen that two experiments are such that the Poisson rule should hold, but that the conditions on  $q$  predict a ratio for the two parameters; the question is whether the data support this ratio. Thus in either case the joint chance of the numbers of occurrences in the two series will be

$$\frac{r^x e^{-r}}{x!} \frac{r'^x e^{-r'}}{x'!}; \quad (1)$$

but on  $q$  we are given  $r/r' = a/(1-a)$ , (2)

and we can introduce  $b$  such that

$$r = ab, \quad r' = (1-a)b; \quad (3)$$

while on  $q'$   $r = \alpha b, \quad r' = (1-\alpha)b,$  (4)

and it now appears that  $\alpha$  must be between 0 and 1. Then

$$P(q, db | H) = f(b) db, \quad P(q' dbd\alpha | H) = f(b) dbd\alpha, \quad (5)$$

$$P(\theta | q, b, H) \propto a^x (1-a)^{x'} b^{x+x'} e^{-b}, \quad (6)$$

$$P(\theta | q', b, \alpha, H) \propto a^x (1-\alpha)^{x'} b^{x+x'} e^{-b}. \quad (7)$$

Hence

$$P(q db | \theta H) \propto f(b) a^x (1-a)^{x'} b^{x+x'} e^{-b} db, \quad (8)$$

$$P(q' dbd\alpha | \theta H) \propto f(b) \alpha^x (1-\alpha)^{x'} b^{x+x'} e^{-b} dbd\alpha. \quad (9)$$

Integration with regard to  $b$  gives the same factor in both cases, and

$$\frac{1}{K} = \int_0^1 \alpha^x (1-\alpha)^{x'} d\alpha \div a^x (1-a)^{x'}, \quad (10)$$

$$K = \frac{(x+x'+1)!}{x! x'!} a^x (1-a)^{x'}. \quad (11)$$

This is the same result as 5.14 (9), but does not depend on the sampling theory of the Poisson rule. It would have several applications where this rule arises. In the case of radioactivity, if  $n$  is the number of atoms in a specimen, and the chance that a given atom will break up in time  $dt$  is  $\lambda dt$ , the expectation of the number in time  $t$  is  $n\lambda t$ . Here  $n$  would be fixed by the mass of the specimen and the atomic weights, and  $t$  by the experimental conditions, while  $\lambda$  is to be found. The need for a significance test would arise if there was a question whether high pressure, temperature, or cosmic rays affected  $\lambda$ . The experiments might not involve the same values of  $n$  and  $t$ , but the expectations, on hypothesis  $q$ , that there is no effect would be in the known ratio  $nt/n't'$ . The test would therefore be given by

$$K = \frac{(x+x'+1)!}{x! x'!} \frac{(nt)^x (n't')^{x'}}{(nt+n't')^{x+x'}}.$$

In the Aitken dust counter, a question might be whether two samples of air are equally dusty. If the same apparatus is used to test both,  $a = \frac{1}{2}$ ; if not,  $a/(1-a)$  is the ratio of the volumes of the samples taken.

Again, two specimens of rock might be compared to see if they are equally radioactive,  $\alpha$ -particle counts being the data. The masses  $m, m'$  of the specimens and the times  $t, t'$  of the experiments would not in general be the same; the expectations of the numbers of disintegrations on the hypothesis that U and Th constitute the same fractions of the specimens will be in the ratio  $mt/m't'$ . This question would seldom arise in practice, since it is highly exceptional for two rocks to have the same radioactivity, but it might arise if there was such a question for two specimens from the same dike.

**5.2. Test of whether the true value in the normal law is zero : standard error originally unknown.** If  $\sigma$  is the standard error and  $\lambda$  the true value,  $\lambda$  is 0 on  $q$ . We want a suitable form for its prior probability on  $q'$ . From considerations of similarity it must depend on  $\sigma$ , since there is nothing in the problem except  $\sigma$  to give a scale for  $\lambda$ . Then we should take

$$P(q' d\sigma | H) \propto \frac{d\sigma}{\sigma}, \quad (1)$$

$$P(q' d\sigma d\lambda | H) \propto f\left(\frac{\lambda}{\sigma}\right) \frac{d\sigma}{\sigma} \frac{d\lambda}{\sigma}, \quad (2)$$

where 
$$\int_{-\infty}^{\infty} f\left(\frac{\lambda}{\sigma}\right) \frac{d\lambda}{\sigma} = 1. \quad (3)$$

If there are  $n$  observations

$$P(\theta | q, \sigma, H) \propto \sigma^{-n} \exp\left\{-\frac{n}{2\sigma^2}(\bar{x}^2 + s'^2)\right\}, \quad (4)$$

$$P(\theta | q', \sigma, \lambda, H) \propto \sigma^{-n} \exp\left[-\frac{n}{2\sigma^2}\{(\bar{x} - \lambda)^2 + s'^2\}\right]. \quad (5)$$

Then

$$P(q d\sigma | \theta H) \propto \sigma^{-n-1} \exp\left\{-\frac{n}{2\sigma^2}(\bar{x}^2 + s'^2)\right\} d\sigma, \quad (6)$$

$$P(q' d\sigma d\lambda | \theta H) \propto f\left(\frac{\lambda}{\sigma}\right) \sigma^{-n-2} \exp\left[-\frac{n}{2\sigma^2}\{(\bar{x} - \lambda)^2 + s'^2\}\right] d\sigma d\lambda. \quad (7)$$

We should expect that for  $n = 1$  no decision would be reached in the absence of previous information about  $\sigma$  and  $\lambda$ , since the departure of a single measure from zero could be interpreted equally well as a random error or as a departure of  $\lambda$  from zero. We should also expect that for  $n \geq 2$ ,  $K$  would be 0 if  $s' = 0$ ,  $\bar{x} \neq 0$ ; for exact agreement of even two observations would be interpreted as an indication that  $\sigma = 0$  and therefore  $\lambda = \bar{x} \neq 0$ .

If  $s' = 0$ ,  $\bar{x} \neq 0$ , take  $\bar{x}$  positive, and put

$$\sigma = \bar{x}/\tau, \quad \lambda = \sigma v = \bar{x}v/\tau. \quad (8)$$

Then

$$P(q | \theta H) \propto \int_0^{\infty} \left(\frac{\tau}{\bar{x}}\right)^n \exp(-\tfrac{1}{2}n\tau^2) \frac{d\tau}{\tau}, \quad (9)$$

$$P(q' | \theta H) \propto \int_0^{\infty} \frac{d\tau}{\tau} \int_{-\infty}^{\infty} \left(\frac{\tau}{\bar{x}}\right)^n f(v) \exp\{-\tfrac{1}{2}n(v - \tau)^2\} dv. \quad (10)$$

(9) converges for all  $n \geq 1$ . If  $n = 1$  and  $f(v)$  is any even function,

$$\begin{aligned} P(q' | \theta H) &\propto \frac{1}{\bar{x}} \int_0^\infty d\tau \int_0^\infty f(v) [\exp\{-\frac{1}{2}(v-\tau)^2\} + \exp\{-\frac{1}{2}(v+\tau)^2\}] dv \\ &= \frac{1}{\bar{x}} \int_{-\infty}^\infty d\tau \int_0^\infty f(v) \exp\{-\frac{1}{2}(v-\tau)^2\} dv \\ &= \frac{\sqrt{(2\pi)}}{\bar{x}} \int_0^\infty f(v) dv = \frac{1}{2} \frac{\sqrt{(2\pi)}}{\bar{x}}. \end{aligned} \quad (11)$$

Also from (9)  $P(q | \theta H) \propto \frac{1}{2} \frac{\sqrt{(2\pi)}}{\bar{x}}, \quad (12)$

and therefore  $K = 1$ . Hence the condition that one observation shall give an indecisive result is satisfied if  $f(v)$  is any even function with integral 1.

If  $n \geq 2$ , the condition that  $K = 0$  for  $s' = 0$ ,  $\bar{x} \neq 0$  is equivalent to the condition that (10) shall diverge. For  $v$  large and positive

$$\int_0^\infty \tau^n \exp\{-\frac{1}{2}n(v-\tau)^2\} \frac{d\tau}{\tau} \sim Nv^{n-1}, \quad (13)$$

where  $N$  is a function of  $n$ . This integral is bounded for small  $v$ . For  $v$  negative it is exponentially small but positive. Hence (10) diverges if and only if

$$\int_0^\infty f(v)v^{n-1} dv \quad (14)$$

diverges. The simplest function satisfying this condition for  $n > 1$  and also satisfying (3) is

$$f(v) = \frac{1}{\pi(1+v^2)}. \quad (15)$$

Corresponding to this and (2)

$$P(d\lambda | q'\sigma H) = \frac{1}{\pi(1+\lambda^2/\sigma^2)} \frac{d\lambda}{\sigma}. \quad (16)$$

In the first edition of this book I used as a parameter a quantity  $\sigma'$ , which would in the present notation be  $(\sigma^2 + \lambda^2)^{1/2}$ , and would have the property that on any set of observations its maximum likelihood estimate would be the same whether  $\lambda$  is assumed zero or not. Then the prior probability of  $\lambda$  was taken uniform with respect to  $\sigma'$ ; hence

$$P(d\sigma' d\lambda | q'H) \propto \frac{d\sigma'}{\sigma'} \frac{d\lambda}{2\sigma'} = \frac{\sigma d\sigma d\lambda}{2(\sigma^2 + \lambda^2)^{3/2}}. \quad (17)$$



This does not satisfy (14) for  $n = 2$ , as was first found in a detailed numerical investigation, which showed that, for  $n = 2$ ,  $K$  could never be less than 0.47 however closely the observations agreed.†

It may be remarked that many physicists totally reject the usual theory of errors on the ground that systematic errors are always present and are not reduced by taking the mean of a large number of observations. They would maintain (1) that the mean of a large number of observations made in the same way is not necessarily better than one observation, and the only use of making more than one observation is to check gross mistakes; (2) that the weighted mean of several series of observations is worse than the value given by the best series. It has been rejected as inconsistent with the theory of probability, but this rejection is associated with the belief that the normal law is the only law of probability. The belief of the old-fashioned physicist can in fact be completely formalized. If the law of error for one observation is a Cauchy law about a constant, then the mean of any number of observations follows exactly the same law, and his condition (1) is satisfied. If, irrespective of the random variation within each series, the location parameter for each set has a departure from the true value with a probability law given by (16), then the mean of the location parameters has a probability distribution of the same form with a scale parameter equal to the mean of the separate  $\sigma$ , and therefore not less than the smallest  $\sigma$ . Thus condition (2) is also satisfied.

On the other hand, detailed study of errors of observation usually shows that they are far from following the Cauchy law; the normal law is nearer, and averages fluctuate less than the Cauchy law would indicate. Also there are plenty of cases where estimates made by different methods have agreed as well as would be expected on the hypothesis that the normal law of error holds and that there are no systematic errors. The belief of the old-fashioned physicist must in fact be regarded as a serious hypothesis, or pair of hypotheses, capable of being sufficiently clearly stated to be tested and therefore deserving test, according to our rule of 1.1 (5). But actual test shows that they are not in general true. We do, however, often find discrepancies. We provide for these by taking prior probability  $\frac{1}{2}$  for no real difference, and  $\frac{1}{2}$  for a real difference, and distributing the latter over possible values of the difference in such a way that if it is not zero it can always be detected and asserted with confidence given sufficient observations. The dependence on the standard error indicated in (16) may be regarded

† *Proc. Roy. Soc. A*, 180, 1942, 256-68.

as an expression of the fact that special care in reducing the random error will usually be associated with special care in eliminating systematic errors. The astronomical case is a special one, since random errors have already been reduced as far as they can for most types of observation, and progress has long depended mainly on eliminating systematic errors. We therefore in our rule of procedure reject the Cauchy law for the random variation about the mean. We use it for systematic differences except that we allow a non-zero fraction, usually  $\frac{1}{2}$ , of the total prior probability to be concentrated at zero difference.

The old-fashioned physicist's view is therefore not nonsensical. It consists of two parts, both of which can be clearly stated, but the first part is wrong and the second exaggerated. When the second part is cleared of exaggeration it leads to a valuable working rule with the properties that we require.

An asymptotic form is easily found for  $K$ , when  $n$  is large. In (7) the large values of the integrand, for given  $\sigma$ , are in a range of order  $\lambda = \bar{x} \pm O(\sigma/\sqrt{n})$ . In such a range  $f(\lambda/\sigma)$  varies little from its value at  $\lambda = \bar{x}$ . Hence we can perform the integration with regard to  $\lambda$  approximately:

$$P(q | \theta H) \propto \int_0^\infty \sigma^{-n-1} \exp\left\{-\frac{n}{2\sigma^2}(\bar{x}^2 + s'^2)\right\} d\sigma, \quad (18)$$

$$P(q' | \theta H) \propto \int_0^\infty \frac{\sqrt{(2\pi)}}{\pi(1 + \bar{x}^2/\sigma^2)\sqrt{n}} \sigma^{-n-1} \exp\left(-\frac{ns'^2}{2\sigma^2}\right) d\sigma. \quad (19)$$

Again, the integrals are of the same form except for the factor in  $\bar{x}/\sigma$ , which varies slowly. The large values of the second integrand are near  $\sigma = s'$ . Substituting this value in the slowly varying factor and suppressing a factor that is the same for both integrals we have

$$P(q | \theta H) \propto (s'^2 + \bar{x}^2)^{-1/2n}, \quad (20)$$

$$P(q' | \theta H) \propto \frac{1}{\pi\sqrt{n}} \left(\frac{2\pi}{n}\right)^{-1/2} \frac{1}{1 + \bar{x}^2/s'^2} s'^{-n}, \quad (21)$$

$$K \sim \sqrt{\left(\frac{\pi n}{2}\right)} \left(1 + \frac{\bar{x}^2}{s'^2}\right)^{-1/2n+1}. \quad (22)$$

The error of the approximations is of the order of  $1/n$  of the whole expression. In terms of

$$t = \sqrt{(n-1)}\bar{x}/s', \quad \nu = n-1, \quad (23)$$

$$K \sim \sqrt{\left(\frac{\pi\nu}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{-1/2\nu+1/2}. \quad (24)$$

The corresponding formula given in the first edition of this book was

$$K \sim \sqrt{\left(\frac{2\nu}{\pi}\right)\left(1 + \frac{t^2}{\nu}\right)^{-1/2\nu+1}}. \quad (25)$$

The new value is larger for  $t$  small and smaller for  $t$  large. We may say that the present test is a little more sensitive.

If  $K$  is very small, so that it is practically certain that  $\lambda$  is not zero, the posterior probability of  $\sigma$  and  $\lambda$  is nearly proportional to

$$P(q' d\lambda d\sigma | \theta H).$$

Comparing (7) with 3.41 (2) we see that the posterior probability is nearly the same as in the estimation problem, being obtained to this accuracy by changing  $\nu$  to  $\nu-1$ .

The behaviour of  $K$  is seen most easily by considering the case when  $\nu$  is large enough for the  $t$  factor to be replaced by  $\exp(-\frac{1}{2}t^2)$ . When  $t = 2$  this is 0.135: when  $t = 3$  it is 0.011. In the former case  $K = 1$  when  $\nu$  is about 30; in the latter  $K = 1$  when  $\nu$  is about 5,000. The variation of  $K$  with  $t$  is much more important than the variation with  $\nu$ ; in fact, for given  $K$ ,  $t$  increases like  $(\log \nu)^{1/2}$ , which is a very slow increase. We may say that if  $t > 3$ ,  $K$  will be less than 1, and the introduction of the new parameter will be supported, for any number of observations that ordinarily occurs. If  $t = 2$ ,  $K$  will be greater than 1 if  $\nu > 30$ , and again for small values of  $\nu$ ; in the case of  $\nu = 2$  and  $t = 2$  the formula (8) makes  $K$  nearly 1, though the accuracy of the approximation is not to be trusted when  $\nu$  is so small. Without elaborate calculation we can then say that values of  $t$  less than 2 will in most cases be regarded as confirming the null hypothesis; values greater than 3 will usually be taken as an indication that the true value is not zero.

The fact that when  $K$  is small the posterior probability of  $\sigma$  and  $\alpha$  is almost the same as in the estimation problem is an indication that we are working on the right lines. There would be no inconsistency in taking  $f(v) \propto e^{-kv^2}$ , where  $k$  is some positive constant, but we have already seen that if we did so  $K$  would never be less than some positive function of  $n$  however closely the observations agreed among themselves. Similarly the posterior probability of  $\sigma$  and  $\alpha$ , even if all the observations agreed exactly, would be the same as if there was an additional observation of positive weight at  $x = 0$ . In cases where the null hypothesis is rejected we should never be led to the conclusion that the standard error was near  $s$  however closely the observations might agree. The chief advantage of the form that we have chosen is that in any significance test it leads to the conclusion that if the null hypothesis

has a small posterior probability, the posterior probability of the parameters is nearly the same as in the estimation problem. Some difference remains, but it is only a trace.

It is also possible to reduce  $1/K$  exactly to a single integral. (18) is exactly

$$P(q | \theta H) \propto \frac{2^{1/2n-1}(\frac{1}{2}n-1)!}{\{n(\bar{x}^2+s'^2)\}^{1/2n}}. \quad (26)$$

From (7) and (15), with

$$\lambda = \sigma v; \quad \frac{n}{2\sigma^2}(\bar{x}^2+s'^2) = u; \quad (27)$$

$$\begin{aligned} P(q' | \theta H) &\propto \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{dv}{1+v^2} \int_0^{\infty} \left( \frac{2u}{n(\bar{x}^2+s'^2)} \right)^{1/2n} \times \\ &\quad \times \exp \left\{ -u + nv\bar{x} \left( \frac{2u}{n(\bar{x}^2+s'^2)} \right)^{1/2} - \frac{nv^2}{2} \right\} \frac{du}{2u} \\ &= \frac{1}{\pi} \frac{2^{1/2n-1}}{\{n(\bar{x}^2+s'^2)\}^{1/2n}} \int_{-\infty}^{\infty} e^{-1/2nv^2} \frac{dv}{1+v^2} \int_0^{\infty} e^{-u} u^{1/2n-1} \times \\ &\quad \times \left\{ 1 + \sum \left( \frac{2nu}{\bar{x}^2+s'^2} \right)^{1/2r} \frac{(v\bar{x})^r}{r!} \right\} du. \end{aligned} \quad (28)$$

Integrate term by term; odd powers of  $v$  contribute nothing to the double integral; and we have

$$\begin{aligned} P(q' | \theta H) &\propto \frac{2^{1/2n-1}(\frac{1}{2}n-1)!}{\pi\{n(\bar{x}^2+s'^2)\}^{1/2n}} \int_{-\infty}^{\infty} \frac{e^{-1/2nv^2} dv}{1+v^2} \times \\ &\quad \times \left\{ 1 + \sum \left( \frac{2n}{\bar{x}^2+s'^2} \right)^m \frac{(v\bar{x})^{2m}}{(2m)!} \frac{1}{2} n \dots (\frac{1}{2}n+m-1) \right\}, \end{aligned} \quad (29)$$

$$\begin{aligned} \frac{1}{K} &= \frac{2}{\pi} \int_0^{\infty} \frac{e^{-1/2nv^2} dv}{1+v^2} \left[ 1 + \sum \frac{\frac{1}{2}n \dots (\frac{1}{2}n+m-1)}{m! \frac{1}{2} \dots (m-\frac{1}{2})} \left\{ \frac{nv^2\bar{x}^2}{2(\bar{x}^2+s'^2)} \right\}^m \right] \\ &= \frac{2}{\pi} \int_0^{\infty} {}_1F_1 \left( \frac{1}{2}n, \frac{1}{2}, \frac{nv^2\bar{x}^2}{2(\bar{x}^2+s'^2)} \right) \frac{e^{-1/2nv^2}}{1+v^2} dv, \end{aligned} \quad (30)$$

where  ${}_1F_1(\alpha, \gamma, x)$  denotes the confluent hypergeometric function

$$1 + \frac{\alpha x}{\gamma} + \frac{\alpha(\alpha+1)x^2}{2!\gamma(\gamma+1)} + \dots \quad (31)$$

By a known identity†

$${}_1F_1(\alpha, \gamma, x) = e^x {}_1F_1(\gamma-\alpha, \gamma, -x); \quad (32)$$

† H. and B. S. Jeffreys, *Methods of Mathematical Physics*, 1946, 576.

hence an alternative form of  $1/K$  is

$$\frac{1}{K} = \frac{2}{\pi} \int_0^{\infty} {}_1F_1\left\{\frac{1}{2} - \frac{1}{2}n, \frac{1}{2}, -\frac{nv^2\bar{x}^2}{2(\bar{x}^2 + s'^2)}\right\} \exp\left\{-\frac{ns'^2v^2}{2(\bar{x}^2 + s'^2)}\right\} \frac{dv}{1+v^2}. \quad (33)$$

**5.21. Test of whether a true value is zero :  $\sigma$  taken as known.** Since (16) is taken to hold for all  $\sigma$ , we can use it when  $\sigma$  is already known; then

$$P(q | \theta H) \propto \exp\left(-\frac{n\bar{x}^2}{2\sigma^2}\right), \quad (34)$$

$$\begin{aligned} P(q' | \theta H) &\propto \frac{1}{\pi\sigma} \int_{-\infty}^{\infty} \exp\left\{-\frac{n}{2\sigma^2}(\bar{x} - \lambda)^2\right\} \frac{d\lambda}{1 + \lambda^2/\sigma^2} \\ &\doteq \sqrt{\left(\frac{2}{\pi n}\right)} \frac{1}{1 + \bar{x}^2/\sigma^2}, \end{aligned} \quad (35)$$

$$K \sim \sqrt{\left(\frac{\pi n}{2}\right)} \left(1 + \frac{\bar{x}^2}{\sigma^2}\right) \exp\left(-\frac{n\bar{x}^2}{2\sigma^2}\right). \quad (36)$$

The method used in the first edition failed to cover this case, but there are many applications where the standard error is so well known from collateral evidence that it can be taken as known.

**5.3. Generalization by invariance theory.** We have seen that for the normal law a satisfactory form of the prior probability is

$$P(d\lambda | q'\sigma H) = \frac{d\lambda}{\pi\sigma(1 + \lambda^2/\sigma^2)}. \quad (1)$$

Now both  $I_2$  and  $J$  of 3.9 (14), (15), when  $\zeta = 0$ , are functions of  $\lambda/\sigma$ ; in fact

$$I_2 = 2\left\{1 - \exp\left(-\frac{1}{8} \frac{\lambda^2}{\sigma^2}\right)\right\}, \quad J = \frac{\lambda^2}{\sigma^2}. \quad (2)$$

$$\text{Then } \frac{d\lambda}{\pi\sigma(1 + \lambda^2/\sigma^2)} = \frac{1}{\pi} d \tan^{-1}\{-8 \log(1 - \frac{1}{2}I_2)\}^{1/2} = \frac{1}{\pi} d \tan^{-1}J^{1/2}, \quad (3)$$

where the square roots are taken to have the same sign as  $\lambda/\sigma$ . The relation to  $J$  is much simpler than the relation to  $I_2$ .

We could therefore make it a general rule in significance tests to express the new parameter in terms of  $I_2$  or  $J$  calculated for comparison of the null hypothesis with the alternative hypothesis, and use prior probabilities of either as given by (3). If the inverse tangents do not range from  $-\frac{1}{2}\pi$  to  $\frac{1}{2}\pi$ , as in cases where the new parameter can take only one sign, correcting factors will be needed. We therefore have possible general rules for significance tests. These rules, however,

disagree with those that we have used in problems of sampling, and our first task must be to see whether they will give satisfactory solutions in those cases.

For the comparison of two sets of chances

$$\left\{ \begin{array}{cc} \alpha\beta & \alpha(1-\beta) \\ (1-\alpha)\beta & (1-\alpha)(1-\beta) \end{array} \right\}, \quad \left\{ \begin{array}{cc} \alpha\beta+\gamma & \alpha(1-\beta)-\gamma \\ (1-\alpha)\beta-\gamma & (1-\alpha)(1-\beta)+\gamma \end{array} \right\} \quad (4)$$

we find  $J_1 = \gamma \log \frac{(\alpha\beta+\gamma)\{(1-\alpha)(1-\beta)+\gamma\}}{\{\alpha(1-\beta)-\gamma\}\{(1-\alpha)\beta-\gamma\}}. \quad (5)$

This would, by the rule just suggested, be suitable to give a prior probability distribution for  $\gamma$  in a contingency problem. Suppose, on the other hand, that we take a sample of  $\phi$ 's and  $\sim\phi$ 's, of given numbers  $n_1, n_2$ , from the class. The chances of  $\psi$  and  $\sim\psi$ , given  $\phi$  and  $\sim\phi$  respectively, will be ( $\gamma$  being 0 on  $q$ )

$$\left. \begin{array}{cc} (\beta+\gamma/\alpha, & 1-\beta-\gamma/\alpha), \\ \{\beta-\gamma/(1-\alpha), & 1-\beta+\gamma/(1-\alpha)\}. \end{array} \right\} \quad (6)$$

Comparing these two pairs of chances we find

$$J_2 = \frac{\gamma}{\alpha(1-\alpha)} \log \frac{(\alpha\beta+\gamma)\{(1-\alpha)(1-\beta)+\gamma\}}{\{\alpha(1-\beta)-\gamma\}\{(1-\alpha)\beta-\gamma\}} = \frac{J_1}{\alpha(1-\alpha)}. \quad (7)$$

If we took samples of  $\psi$ 's and  $\sim\psi$ 's and counted the  $\phi$ 's and  $\sim\phi$ 's, we should get similarly

$$J_3 = \frac{J_1}{\beta(1-\beta)}. \quad (8)$$

To satisfy the condition that the significance test, for given sampling numbers, should be nearly independent of the conditions of sampling, the prior probability of  $\gamma$ , given  $\alpha$  and  $\beta$ , should be the same in all cases. Hence we cannot simply use  $J$  universally. But we can define a  $J$  that would be equal, for given  $\gamma$ , in the three cases, and with the proper properties of symmetry, by taking either

$$J_1, \quad \alpha(1-\alpha)J_2, \quad \beta(1-\beta)J_3 \quad (9)$$

or 
$$\frac{J_1}{\alpha(1-\alpha)\beta(1-\beta)}, \quad \frac{J_2}{\beta(1-\beta)}, \quad \frac{J_3}{\alpha(1-\alpha)}. \quad (10)$$

The first set are plainly unsatisfactory. For  $J$  tends to infinity at the extreme possible values of  $\gamma$ ; hence if the estimate of  $\gamma$  is a small quantity  $c$  it will lead to

$$K \sim \sqrt{\left(\frac{\pi N}{2}\right)} \exp(-\tfrac{1}{2}\chi^2),$$

where  $N$  is the sum of the sample numbers. This conflicts with the rule

of 5.03 that the outside factor should be of order  $(x+y)^{1/2}$ , where  $x+y$  is the smallest of the row and column totals. On the other hand, the second set are consistent with this rule.

A minor objection is that two pairs of chances expressed in the form (6) do not suffice to determine  $\alpha$ ,  $\beta$ , and  $\gamma$ , and there is some indeterminacy as to what we shall take for  $\beta$  in (10). But so long as  $\gamma$  is small it will make little difference what value between  $\beta+\gamma/\alpha$  and  $\beta-\gamma/(1-\alpha)$  we choose.

$I_2$  is much less satisfactory in this problem. There is no simple exact relation between the values of  $I_2$  in the three comparisons made. Also  $I_2$  takes finite values (not 2) for the extreme possible values of  $\gamma$  if neither  $\alpha$  nor  $\beta$  is 0 or 1. It appears therefore that  $I_2$  cannot be made to satisfy the conditions by any linear transformation. In view of the greater complexity of the expression in  $I_2$  in (3) than of that in  $J$ , it appears unnecessary to pay further attention to  $I_2$  at present.

An objection to  $J$ , even in the modified form, is that if the suggested value of a chance is 1 comparison with any other value gives  $J$  infinite. Consequently the rule based on  $J$  in (3) would concentrate the whole of the prior probability of the chance in the value 1 on the alternative hypothesis, which thereby becomes identical with the null hypothesis. Of course a single exception to the rule would disprove the null hypothesis deductively in such a case, but nevertheless the situation is less satisfactory than in the analysis given in 5.1. It might even be said that the use of anything as complicated as  $J$  in so simple a problem as the testing of a suggested chance is enough to condemn it.

It appears to be worth recording the asymptotic forms given by (10) in the problems of 5.1. We find without much difficulty

$$K \sim \{\pi(x+y)pp'\}^{1/2} \exp(-\frac{1}{2}\chi^2) \quad \text{for 5.1 (9),}$$

$$K \sim \left\{ \frac{\pi(x+y)(x+x')(x'+y')(y+y')}{2N^3} \right\}^{1/2} \exp(-\frac{1}{2}\chi^2)$$

for 5.11 (18), 5.12 (10), 5.13 (3).

An evaluation of  $K$  has also been made for the problem of 5.11, using the estimation prior probabilities given by the invariance rules. It was again of the order of  $N^{1/2}$ . These attempts at using the invariance theory in sampling problems, therefore, confirm the suggestion of 3.9 (p. 163) that there is nothing to be gained by attempting to ensure general invariance for transformation of chances; uniform distribution within the permitted intervals is more satisfactory, as far as can be seen at present. We shall, however, use the rule based on  $J$  in the more

complicated cases where there is no obvious suggestion from more elementary ones.

**5.31. General approximate forms.** We see from 3.9(3) that if a new parameter  $\alpha$  is small,

$$J \doteq g_{\alpha\alpha}\alpha^2, \quad (1)$$

and if  $\alpha$  can take either sign, the range of possible values being such that  $J$  can tend to infinity for variations of  $\alpha$  in either direction,

$$P(d\alpha | q'H) = \frac{|dJ^{1/2}|}{\pi(1+J)} \doteq g_{\alpha\alpha}^{1/2} \frac{d\alpha}{\pi} \quad (2)$$

for  $\alpha$  small. If  $n$  observations yield an estimate  $\alpha = a$ , where  $na^3$  can be neglected,

$$\log L \doteq \frac{1}{2}ng_{\alpha\alpha}(\alpha-a)^2. \quad (3)$$

Hence in 5.0(4) we can put

$$f(\alpha) = g_{\alpha\alpha}^{1/2}/\pi; \quad s = 1/(ng_{\alpha\alpha})^{1/2}, \quad (4)$$

and then

$$K \sim \left(\frac{\pi n}{2}\right)^{1/2} \exp\left(-\frac{a^2}{2s^2}\right). \quad (5)$$

If  $\alpha$  can take values only on one side of 0, (2) must be doubled for  $\alpha$  on that side, and if  $a$  also is on that side the value of  $K$  given by (5) will be approximately halved. If  $a$  is on the other side of 0, the approximate form fails; we shall see that  $K$  may then be of order  $n$  instead of  $n^{1/2}$ .

The approximate form will be adequate for practical purposes in the majority of problems. Closer approximations are needed when  $n$  is small: for instance, in problems concerned with the normal law the need to estimate the standard error also from the same set of observations may make an appreciable difference. But if  $n$  is more than 50 or so (5) can be used as it stands without risk of serious mistakes.

#### 5.4. Other tests related to the normal law.

**5.41. Test of whether two true values are equal, standard errors supposed the same.** This problem will arise when two sets of observations made by the same method are used to detect a new parameter by their difference. According to the rule that we are adopting, any series of observations is suspected of being subject to disturbance until there is reason to the contrary. When we are comparing two series, therefore, we are really considering four hypotheses, not two as in the test for agreement of a location parameter with zero; for neither may be disturbed, or either or both may. We continue to denote the hypothesis that both location parameters are  $\lambda$  by  $q$ , but  $q'$



is broken up into three, which we shall denote by  $q_1$ ,  $q_2$ ,  $q_{12}$ . With an obvious notation we therefore take

$$P(q \, d\sigma d\lambda | H) \propto d\sigma d\lambda / \sigma, \quad (1)$$

$$P(q_1 \, d\sigma d\lambda d\lambda_1 | H) \propto \frac{1}{\pi} d\sigma d\lambda \frac{d\lambda_1}{\sigma^2 + (\lambda_1 - \lambda)^2}, \quad (2)$$

$$P(q_2 \, d\sigma d\lambda d\lambda_2 | H) \propto \frac{1}{\pi} \frac{d\sigma d\lambda d\lambda_2}{\sigma^2 + (\lambda_2 - \lambda)^2}, \quad (3)$$

$$P(q_{12} \, d\sigma d\lambda d\lambda_1 d\lambda_2 | H) \propto \frac{1}{\pi^2} \frac{\sigma \, d\sigma d\lambda d\lambda_1 d\lambda_2}{\{\sigma^2 + (\lambda_1 - \lambda)^2\} \{\sigma^2 + (\lambda_2 - \lambda)^2\}}. \quad (4)$$

On  $q_1$ ,  $\lambda_2 = \lambda$ ; on  $q_2$ ,  $\lambda_1 = \lambda$ . On  $q_{12}$ , since  $\lambda$  does not appear explicitly in the likelihood, we can integrate with regard to it immediately:

$$P(q_{12} \, d\sigma d\lambda_1 d\lambda_2 | H) \propto \frac{2}{\pi} \frac{d\sigma d\lambda_1 d\lambda_2}{4\sigma^2 + (\lambda_1 - \lambda_2)^2}. \quad (5)$$

Also

$$P(\theta | \sigma \lambda_1 \lambda_2 H) \propto \sigma^{-n_1 - n_2} \exp \left\{ -\frac{n_1}{2\sigma^2} (\bar{x}_1 - \lambda_1)^2 - \frac{n_2}{2\sigma^2} (\bar{x}_2 - \lambda_2)^2 - \frac{n_1 s_1'^2 + n_2 s_2'^2}{2\sigma^2} \right\}. \quad (6)$$

Put  $\nu = n_1 + n_2 - 2$ ;  $\nu s^2 = n_1 s_1'^2 + n_2 s_2'^2$ .  
Then (7)

$$P(q \, d\sigma d\lambda | \theta H) \propto \sigma^{-n_1 - n_2} \exp \left\{ -\frac{n_1}{2\sigma^2} (\bar{x}_1 - \lambda)^2 - \frac{n_2}{2\sigma^2} (\bar{x}_2 - \lambda)^2 - \frac{\nu s^2}{2\sigma^2} \right\} \frac{d\sigma d\lambda}{\sigma} \quad (8)$$

with corresponding equations for  $q_1$ ,  $q_2$ , and  $q_{12}$ . It is easy to verify that the posterior probabilities of all four hypotheses are equal if  $n_1 = 1$ ,  $n_2 = 0$  or if  $n_1 = n_2 = 1$ , as we should expect. If  $n_1$  and  $n_2$  are large we find, approximately,

$$\begin{aligned} P(q | \theta H) : P(q_1 | \theta H) : P(q_2 | \theta H) : P(q_{12} | \theta H) \\ = \left( \frac{\pi}{2} \frac{n_1 n_2}{n_1 + n_2} \right)^{1/2} \left\{ 1 + \frac{(\bar{x}_1 - \bar{x}_2)^2}{s^2} \right\} \left\{ 1 + \frac{n_1 n_2}{n_1 + n_2} \frac{(\bar{x}_1 - \bar{x}_2)^2}{s^2} \right\}^{-1/2(n_1 + n_2 - 1)} \\ : 1 : 1 : \frac{1}{2} \frac{s^2 + (\bar{x}_1 - \bar{x}_2)^2}{s^2 + \frac{1}{4}(\bar{x}_1 - \bar{x}_2)^2}. \end{aligned} \quad (9)$$

The equality of the second and third of these numbers is of course exact. The last ranges from  $\frac{1}{2}$  to 2 as  $|\bar{x}_1 - \bar{x}_2|/s$  increases from 0 to infinity. Thus the test never expresses any decision between  $q_1$  and  $q_2$ , as we should expect, and never expresses one for  $q_{12}$  against  $q_1$  v  $q_2$ . It expresses a slight preference for  $q_{12}$  against  $q_1$  or  $q_2$  separately if  $|\bar{x}_1 - \bar{x}_2|/s > \sqrt{2}$ . But there is so little to choose between the alternatives that we may

as well combine them. If  $|\bar{x}_1 - \bar{x}_2|/s$  is small, as it usually will be, we can write

$$\frac{P(q|\theta H)}{P(q_1 \vee q_2 \vee q_{12})} \sim \frac{2\left(\pi \frac{n_1 n_2}{2n_1 + n_2}\right)^{1/2}}{5} \left\{1 + \frac{n_1 n_2}{n_1 + n_2} \frac{(\bar{x}_1 - \bar{x}_2)^2}{s^2}\right\}^{-1/2(n_1 + n_2 - 1)}. \quad (10)$$

Expressing the standard errors of  $x_1$  and  $x_2$  in the usual way,

$$s_{x_1}^2 = s^2/n_1, \quad s_{x_2}^2 = s^2/n_2, \quad (11)$$

$$s_{x_1 - x_2}^2 = \frac{n_1 + n_2}{n_1 n_2} s^2, \quad t = (\bar{x}_1 - \bar{x}_2)/s_{x_1 - x_2}, \quad (12)$$

we can write (10) as

$$\frac{2\left(\pi \frac{n_1 n_2}{2n_1 + n_2}\right)^{1/2}}{5} \left(1 + \frac{t^2}{\nu}\right)^{-1/2(\nu+1)}. \quad (13)$$

Decision between the three alternative hypotheses, in cases where this ratio is less than 1, will require additional evidence such as a comparison with a third series of observations.

Where there is strong reason to suppose that the first series gives an estimate of a quantity contained in a theory and is free from systematic error, the alternatives  $q_1$  and  $q_{12}$  do not arise, and the factor  $2/5$  in (13) is unnecessary.

**5.42. Test of whether two location parameters are the same, standard errors not supposed equal.** The method is substantially as in the last section, and leads to the following equations:

$$P(q d\sigma_1 d\sigma_2 | \theta H) \propto \sqrt{\left(\frac{2\pi}{n_1 n_2}\right) \frac{\sigma_1^{-n_1} \sigma_2^{-n_2}}{(\sigma_1^2/n_1 + \sigma_2^2/n_2)}} \times \\ \times \exp\left\{-\frac{n_1 s_1'^2}{2\sigma_1^2} - \frac{n_2 s_2'^2}{2\sigma_2^2} - \frac{(\bar{x}_1 - \bar{x}_2)^2}{2(\sigma_1^2/n_1 + \sigma_2^2/n_2)}\right\} d\sigma_1 d\sigma_2, \quad (1)$$

$$P(q_1 d\sigma_1 d\sigma_2 | \theta H) \propto \frac{2}{\sqrt{(n_1 n_2)}} \frac{\sigma_1^{-n_1+1} \sigma_2^{-n_2}}{\sigma_1^2 + (\bar{x}_1 - \bar{x}_2)^2} \exp\left(-\frac{n_1 s_1'^2}{2\sigma_1^2} - \frac{n_2 s_2'^2}{2\sigma_2^2}\right) d\sigma_1 d\sigma_2, \quad (2)$$

$P(q_2 d\sigma_1 d\sigma_2 | \theta H)$  follows by symmetry,

$$P(q_{12} d\sigma_1 d\sigma_2 | \theta H) \propto \frac{2}{\sqrt{(n_1 n_2)}} \frac{\sigma_1^{-n_1} \sigma_2^{-n_2}}{(\sigma_1 + \sigma_2) \{1 + (\bar{x}_1 - \bar{x}_2)^2/(\sigma_1 + \sigma_2)^2\}} \times \\ \times \exp\left(-\frac{n_1 s_1'^2}{2\sigma_1^2} - \frac{n_2 s_2'^2}{2\sigma_2^2}\right) d\sigma_1 d\sigma_2. \quad (3)$$

The form of the term in  $(\bar{x}_1 - \bar{x}_2)^2$  in (1) makes further approximations awkward for general values of  $\bar{x}_1 - \bar{x}_2$ , but we may appeal to the fact that  $K$  is usually small when the maximum likelihood estimate of a

new parameter is more than about 3 times its apparent standard error. If we have a good approximation when  $|\bar{x}_1 - \bar{x}_2|$  is less than

$$3\sqrt{(s_1'^2/n_1 + s_2'^2/n_2)},$$

it will be useful up to values that make  $K$  small, and as it will be smaller still for larger values we cannot be led into saying that  $K$  is small when it is not. The precise evaluation of  $K$  when it is very small is not important; it makes no difference to our further procedure if we estimate  $K$  as  $10^{-3}$  when it is really  $10^{-2}$ , since we shall adopt the alternative hypothesis in either case. With these remarks we note that if we can replace  $(\sigma_1^2/n_1 + \sigma_2^2/n_2)^{-1}$  by  $A_1/\sigma_1^2 + A_2/\sigma_2^2$ , where  $A_1, A_2$  are chosen so that the functions and their first derivatives are equal when  $\sigma_1 = s_1, \sigma_2 = s_2$ , the exponent in (1) will be sufficiently accurately represented over the whole range where the integrand is not small in comparison with its maximum. This condition is satisfied if we take

$$A_1 = \frac{s_1^4/n_1}{(s_1^2/n_1 + s_2^2/n_2)^2}; \quad A_2 = \frac{s_2^4/n_2}{(s_1^2/n_1 + s_2^2/n_2)^2}. \quad (4)$$

Replacing  $\sigma_1$  and  $\sigma_2$  by  $s_1$  and  $s_2$  in factors raised to small powers and dropping common factors we find, with  $\nu_1 = n_1 - 1, \nu_2 = n_2 - 1$ ,

$$P(q|\theta H) \propto \sqrt{(\frac{1}{2}\pi)} \frac{1}{(s_1^2/n_1 + s_2^2/n_2)^{1/2}} \left\{ 1 + \frac{s_1^2(\bar{x}_1 - \bar{x}_2)^2/n_1 \nu_1}{(s_1^2/n_1 + s_2^2/n_2)^2} \right\}^{-1/2(\nu_1 - 1)} \times \\ \times \left\{ 1 + \frac{s_2^2(\bar{x}_1 - \bar{x}_2)^2/n_2 \nu_2}{(s_1^2/n_1 + s_2^2/n_2)^2} \right\}^{-1/2(\nu_2 - 1)}, \quad (5)$$

$$P(q_1|\theta H) \propto \frac{s_1}{s_1^2 + (\bar{x}_1 - \bar{x}_2)^2}, \quad (6)$$

$$P(q_2|\theta H) \propto \frac{s_2}{s_2^2 + (\bar{x}_1 - \bar{x}_2)^2}, \quad (7)$$

$$P(q_{12}|\theta H) \propto \frac{s_1 + s_2}{(s_1 + s_2)^2 + (\bar{x}_1 - \bar{x}_2)^2}. \quad (8)$$

There may here be considerable grounds for decision between the alternative hypotheses  $q_1, q_2, q_{12}$ . We recall that  $q_1$  is the hypothesis that the first series is disturbed and not the second, and our approximations contemplate that  $|\bar{x}_1 - \bar{x}_2|$  is small compared with  $s_1$  and  $s_2$ . Then if  $s_1$  is much less than  $s_2$ ,  $P(q_1|\theta H)$  will be much more than  $P(q_2|\theta H)$ , and  $P(q_{12}|\theta H)$  will be slightly less than the latter. That is, subject to the condition that either series of observations is initially regarded as equally likely to be disturbed, the result of many observations will be to indicate that the series with the larger standard deviation is the less likely if the discrepancy is small compared with the

standard errors of one observation. If the approximations remain valid (which has not been investigated) the contrary will hold if the discrepancy between the means is greater than either standard error of one observation.

**5.43. Test of whether a standard error has a suggested value  $\sigma_0$ .** We take the true value to be 0. If the standard error is  $\sigma$ , and

$$\sigma = \sigma_0 e^{\zeta}, \quad (1)$$

$$\text{we have from 3.9 (15)} \quad J = 2 \sinh^2 \zeta \quad (2)$$

$$\text{and} \quad \frac{1}{\pi} d \tan^{-1} J^{1/2} = \frac{\sqrt{2} \cosh \zeta}{\pi \cosh 2\zeta} d\zeta. \quad (3)$$

Then according to 5.3 (3) we should take

$$P(q | H) = \frac{1}{2}; \quad P(q' d\sigma | H) = \frac{1}{\pi\sqrt{2}} \frac{\cosh \zeta}{\cosh 2\zeta} d\zeta. \quad (4)$$

If there are  $n$  observations and the mean square deviation from 0 is  $s^2$ ,

$$P(\theta | qH) \propto \sigma_0^{-n} \exp\left(-\frac{ns^2}{2\sigma_0^2}\right), \quad (5)$$

$$P(\theta | q'H) \propto \sigma^{-n} \exp\left(-\frac{ns^2}{2\sigma^2}\right), \quad (6)$$

$$P(q | \theta H) \propto \sigma_0^{-n} \exp\left(-\frac{ns^2}{2\sigma_0^2}\right), \quad (7)$$

$$P(q' | \theta H) \propto \frac{\sqrt{2}}{\pi} \int_{-\infty}^{\infty} \frac{\cosh \zeta}{\cosh 2\zeta} \sigma^{-n} \exp\left(-\frac{ns^2}{2\sigma^2}\right) d\zeta. \quad (8)$$

The factors with  $n$  in the index have a maximum when  $\sigma = s$ . Put

$$s/\sigma_0 = e^z. \quad (9)$$

For large  $n$  the expression (8) is approximately

$$\frac{\sqrt{2}}{\pi} \frac{\cosh z}{\cosh 2z} s^{-n} \exp\left(-\frac{1}{2}n\right) \sqrt{\left(\frac{\pi}{n}\right)} \quad (10)$$

$$\text{and} \quad K \sim \sqrt{\left(\frac{\pi n}{2}\right)} \frac{\cosh 2z}{\cosh z} e^{nz} \exp\left\{\frac{1}{2}n(1 - e^{2z})\right\}. \quad (11)$$

This is greatest when  $z = 0$  and is then  $\sqrt{(\frac{1}{2}\pi n)}$ .

If instead of using  $J$  we had used  $I_2$  as in 5.3 (3), we should have had instead of the second of (4)

$$P(q' d\sigma | H) = \frac{1}{\pi} \frac{d\zeta}{\cosh 2\zeta} \quad (12)$$

and the first two factors in (11) would be replaced by

$$\frac{1}{2} \sqrt{(\pi n)} \cosh 2z. \quad (13)$$

An exact form of  $1/K$  is

$$\frac{1}{K} = \frac{\sqrt{2}}{\pi} \int_0^{\infty} \frac{u^2+1}{u^4+1} u^n \exp\{\frac{1}{2}nb^2(1-u^2)\} du, \quad (14)$$

where  $\sigma = \sigma_0/u$ ,  $s = \sigma_0 b$ ,  $b = e^z$ . It is seen that this tends to infinity for  $n = 1$  if  $b \rightarrow 0$  or  $b \rightarrow \infty$ . (12) would give for  $n = 1$

$$\frac{1}{K} = \frac{2}{\pi} \int_0^{\infty} \frac{u^2}{u^4+1} \exp\{\frac{1}{2}b^2(1-u^2)\} du, \quad (15)$$

which tends to a finite limit as  $b \rightarrow 0$ . (14) is more satisfactory because it says that one deviation, if small enough, can give strong evidence against  $q$ ; (15) does not. Either gives  $1/K$  large if  $b$  is large.

It has been supposed that all values of  $\sigma$  are admissible on  $q'$ ; the conditions contemplate a theory that predicts a definite standard error  $\sigma_0$ , but we may be ready to accept a standard error either more or less than the predicted value. But where there is a predicted standard error the type of disturbance chiefly to be considered is one that will make the actual one larger, and verification is desirable before the predicted value is accepted. Hence we consider also the case where  $\zeta$  is restricted to be non-negative. The result is to change  $\sqrt{2}$  in (8) to  $2\sqrt{2}$  and make the lower limit 0. The approximations now fall into three types according as  $\zeta = z$  lies well within the range of integration, well outside it, or near 0.

If  $z > 0$  and  $nz^2$  is more than 4 or so, the large values of the integrand on both sides of the maximum lie within the range of integration and the integral is little altered; then the only important change is that  $K$  as given by (11) must be halved.

If  $z = 0$ , only the values on one side of the maximum lie in the range of integration and the integral is halved; this cancels the extra factor 2 and the result is unaltered.

If  $z < 0$  and  $nz^2$  is large, the integrand decreases rapidly from  $\zeta = 0$ . In fact

$$\sigma^{-n} \exp\left(-\frac{ns^2}{2\sigma^2}\right) \doteq \sigma_0^{-n} \exp\left(-\frac{ns^2}{2\sigma_0^2}\right) \exp\{-n(1-e^{-2z})\zeta\} \quad (16)$$

and

$$K \sim \frac{1}{2}\pi n(1-e^{2z}). \quad (17)$$

The factor  $n$  in the last expression instead of the usual  $n^{1/2}$  needs comment. In the usual conditions of a significance test the maximum likelihood solution, here  $\zeta = z$ , or  $\sigma = s$ , is a possible value on  $q'$ . But here we are considering a case where the maximum likelihood solution corresponds to a value of  $\sigma$  that is impossible on  $q'$ , and is less probable

on any value of  $\sigma$  compatible with  $q'$  than on  $q$ . Naturally, therefore, if such a value should occur it would imply unusually strong support for  $q$ . Actually, however, such values will be rare, and if they occur they will not as a rule be accepted as confirming  $q$ , as we shall see later (p. 281).

In the above treatment the true value has been taken as known. If it is unknown (5) and (6) need modification. If we redefine  $s$  as the standard deviation and put  $n-1 = \nu$ , integration with regard to  $\lambda$  will remove a factor  $1/\sigma_0$  from (7) and  $1/\sigma$  from (8). The result will be that  $n$  in (11) and (13) will be replaced by  $\nu$ .

**5.44. Test of agreement of two estimated standard errors.** We shall consider the case where only one of them,  $\sigma_1$ , is possibly disturbed.

$$\text{Put} \quad \sigma_1 = \sigma_2 e^{\zeta}. \quad (1)$$

Then

$$P(q \, d\sigma \mid H) \propto \frac{d\sigma}{\sigma}; \quad P(q' \, d\sigma_1 \, d\sigma_2 \mid H) \propto \frac{\sqrt{2}}{\pi} \frac{\cosh \zeta}{\cosh 2\zeta} d\zeta \frac{d\sigma_2}{\sigma_2}, \quad (2)$$

$$P(\theta \mid q\sigma H) \propto \sigma^{-n_1-n_2} \exp\left(-\frac{n_1 s_1^2 + n_2 s_2^2}{2\sigma^2}\right), \quad (3)$$

$$P(\theta \mid q'\sigma_1 \sigma_2 H) \propto \sigma_1^{-n_1} \sigma_2^{-n_2} \exp\left(-\frac{n_1 s_1^2}{2\sigma_1^2} - \frac{n_2 s_2^2}{2\sigma_2^2}\right), \quad (4)$$

$$P(q \mid \theta H) \propto \int_0^\infty \sigma^{-n_1-n_2} \exp\left(-\frac{n_1 s_1^2 + n_2 s_2^2}{2\sigma^2}\right) \frac{d\sigma}{\sigma}, \quad (5)$$

$$P(q' \mid \theta H) \propto \int_0^\infty \frac{d\sigma_2}{\sigma_2} \int_0^\infty \sigma_1^{-n_1} \sigma_2^{-n_2} \exp\left(-\frac{n_1 s_1^2}{2\sigma_1^2} - \frac{n_2 s_2^2}{2\sigma_2^2}\right) \frac{\sqrt{2}}{\pi} \frac{\cosh \zeta}{\cosh 2\zeta} \frac{d\sigma_1}{\sigma_1}. \quad (6)$$

$$\text{Put} \quad s_1 = s_2 e^{\zeta}. \quad (7)$$

Then

$$P(q' \mid \theta H) \propto \int_0^\infty \frac{d\sigma_2}{\sigma_2} \int_{-\infty}^\infty \sigma_2^{-n_1-n_2} e^{-n_1 \zeta} \exp\left\{-\frac{s_2^2}{2\sigma_2^2} (n_1 e^{2\zeta-2} + n_2)\right\} \times \\ \times \frac{\sqrt{2}}{\pi} \frac{\cosh \zeta}{\cosh 2\zeta} d\zeta, \quad (8)$$

$$\frac{1}{K} = \frac{\sqrt{2}}{\pi} \int_{-\infty}^\infty \frac{\cosh \zeta}{\cosh 2\zeta} e^{-n_1 \zeta} \left(\frac{n_1 e^{2\zeta-2} + n_2}{n_1 e^{2\zeta} + n_2}\right)^{-1/2(n_1+n_2)} d\zeta. \quad (9)$$

The factors with large indices have a maximum when  $\zeta = z$ , and we get approximately

$$K = \left\{\frac{\pi n_1 n_2}{2(n_1+n_2)}\right\}^{1/2} \frac{\cosh 2z}{\cosh z} e^{n_1 z} \left(\frac{n_1+n_2}{n_1 e^{2z} + n_2}\right)^{1/2(n_1+n_2)}. \quad (10)$$

$K$  is unaltered if  $n_1$  and  $n_2$  are interchanged and the sign of  $z$  is reversed.

If, in addition,  $z$  is fairly small, a further approximation gives

$$K \sim \left\{ \frac{\pi n_1 n_2}{2(n_1 + n_2)} \right\}^{1/2} (1 + \frac{3}{2} z^2) \exp \left( - \frac{n_1 n_2 z^2}{n_1 + n_2} \right). \quad (11)$$

If either or both of the standard errors is regarded as possibly disturbed,  $K$  can be adjusted as in 5.41 by multiplication by a factor between  $\frac{1}{2}$  and  $\frac{1}{3}$ . Such conditions might arise when two methods of measurement have a great deal in common, but differ in other features, and it is uncertain which is the better.

The more usual types of case where we wish to compare two standard deviations for consistency are, first, when it is suspected that some additional disturbance has increased the standard error in one set; secondly, when methods have been altered in the hope of reducing the standard error and we want to know whether they have been successful. In the first case we expect  $\zeta$  if not zero to be positive, in the latter negative. We take the former case; then the second of (2) must be multiplied by 2 and the range of  $\zeta$  taken to be from 0 to  $\infty$ . If  $z$  is positive and

$$\frac{n_1 n_2 z^2}{n_1 + n_2} > 2$$

the net result is that  $K$  as given by (10) or (11) should be halved. If  $z$  is negative  $K$  may be large of order  $n_1$  or  $n_2$ .

**5.45. Test of both the standard error and the location parameter.** If the null hypothesis is that two sets of data are derived from the same normal law, we may need to test consistency of both the standard errors and the location parameters. There are cases where we need to arrange the work, when several new parameters are considered, so that the results will be independent of the order in which they are tested. This, I think, is not one of them. The question of consistency of the location parameters is hardly significant until we have some idea of the scale parameters, and if there is serious doubt about whether these are identical it seems nearly obvious that it should be resolved first.

**5.46.** The following example, supplied to me by Professor C. Teodorescu of Timisoara, illustrates the method of 5.42. There was a suggestion that locomotive and wagon tires might be stronger near the edges than in the centre, since they are subjected to more severe working there in the process of manufacture. A number of test pieces were cut, and a tensile test was made on each. The breaking tension,  $R$ , in kilograms weight per mm.<sup>2</sup>, and the percentage extension afterwards,  $A$ , were recorded. In the first place the whole of the data are treated

as two independent series, of 150 observations each. For the edge pieces the mean strength  $R_1$  is found to be 89.59 kg./mm.<sup>2</sup>,  $s_1^2$ , in the corresponding unit, 7.274. For the centre pieces the mean is

$$R_2 = 88.17 \text{ kg./mm.}^2, \quad s_2^2 = 5.619.$$

$$s_1^2/n_1 = 0.04849; \quad s_2^2/n_2 = 0.03746; \quad \bar{x}_1 - \bar{x}_2 = +1.42;$$

$$\begin{aligned} P(q | \theta H) &\propto \sqrt{(\frac{1}{2}\pi)} \frac{1}{\sqrt{(0.08595)}} \left(1 + \frac{0.04849 \times 1.42^2}{149 \times 0.08595^2}\right)^{-149/2} \times \\ &\quad \times \left(1 + \frac{0.03746 \times 1.42^2}{149 \times 0.08595^2}\right)^{-149/2} \\ &= 4.27(1.15727)^{-149/2} = 7.8 \times 10^{-5}, \end{aligned}$$

$$P(q_1 | \theta H) \propto \frac{2.70}{7.27 + 2.02} = 0.29, \quad P(q_2 | \theta H) \propto \frac{2.37}{5.62 + 2.00} = 0.31,$$

$$P(q_{12} | \theta H) \propto \frac{5.07}{25.7 + 2.00} = 0.18,$$

$$\frac{P(q | \theta H)}{P(q_1 \vee q_2 \vee q_{12} | \theta H)} \doteq \frac{7.8 \times 10^{-5}}{0.78} = 10^{-4}.$$

For the extensions the means were  $A_1 = 12.60$  per cent.,  $A_2 = 12.33$  per cent., with  $s_1^2 = 1.505$ ,  $s_2^2 = 1.425$ ; we find similarly

$$P(q | \theta H) \propto 9.0 \times 0.1530 = 1.38,$$

$$P(q_1 | \theta H) \propto 0.78, \quad P(q_2 | \theta H) \propto 0.81, \quad P(q_{12} | \theta H) \propto 0.41,$$

$$\frac{P(q | \theta H)}{P(q_1 \vee q_2 \vee q_{12} | \theta H)} \doteq \frac{1.38}{2.00} = 0.69.$$

Thus there is strong evidence for a systematic difference in the strengths. The result for a difference in the extensions is indecisive.

Since, however, the question asked directly is 'Has the extra working at the edges had a systematic effect?' it may be held that  $q_2$  and  $q_{12}$  do not arise and that we need only consider  $q_1$ . Then for the strengths we find

$$\frac{P(q | \theta H)}{P(q_1 | \theta H)} \doteq \frac{7.8 \times 10^{-5}}{0.29} = 2.7 \times 10^{-4}$$

and for the extensions

$$\frac{P(q | \theta H)}{P(q_1 | \theta H)} \doteq \frac{1.38}{0.78} = 1.8.$$

This way of looking at the data, however, omits an important piece



of information, since the pairs of values for different specimens *from the same tire* were available. There is also a strong possibility of differences between tires; that is why testing was undertaken before comparison of centres and edges. This was so well established that it can be treated as a datum. But then differences between tires will have contributed to the various values of  $s^2$ , without affecting the differences of the means. Hence the above values of  $K$  will be too high considering this additional information. (If this effect was in doubt it could be tested by means of the test for the departure of a correlation coefficient from zero.) A more accurate test can therefore be obtained by treating the differences between values for the same tire as our data, and testing whether they differ significantly from 0. For the differences in  $R$  we find  $s'^2 = 3.790$ , for those in  $A$ ,  $s'^2 = 1.610$ , and we can use the simple formula 5.2 (22). Then for  $R$

$$K \doteq \left(\frac{150\pi}{2}\right)^{1/2} \left(1 + \frac{1.42^2}{3.79}\right)^{-74} = 3 \times 10^{-13}$$

and for  $A$        $K \doteq \left(\frac{150\pi}{2}\right)^{1/2} \left(1 + \frac{0.27^2}{1.6}\right)^{-74} = 0.58.$

The evidence is now overwhelming for a difference in  $R$  and slightly in favour of a difference in  $A$ . This indicates how treatment of a systematic variation as random may obscure other systematic variations by inflation of the standard error; but if comparisons for the same tire had not been available the first test would have been the only one possible. We notice that for  $R$  the variation of the differences between pieces from the same tire is less than the variation of either the centre or the edge pieces separately. For  $A$  it is a little greater; but if the variations were independent we should have expected the mean square variation to be about  $1.495 + 1.416 = 2.91$  instead of the observed 1.61.

The explanation of the much less decisive result for  $A$  even with the more accurate treatment may be that while  $R$  will depend on the least strength of any part of the specimen, the actual process of fracture includes a great deal of continuous flow, and while the stronger material is under a greater stress in the test it may also be relatively less ductile, so that two systematic effects partly cancel.

**5.47. The discovery of argon.** Rayleigh's data† in this investigation refer to the mass of nitrogen obtained from air or by chemical methods,

† *Proc. Roy. Soc.* 53, 1893, 145; 55, 1894, 340-4.

within a given container at standard temperature and pressure. All are in grams.

*From air.*

<i>By hot copper</i>	<i>By hot iron</i>	<i>By ferrous hydrate</i>
2.31035	2.31017	2.31024
26	0986	10
24	1010	28
12	1001	
27		

*By chemical methods.*

<i>Iron and NO</i>	<i>Iron and N<sub>2</sub>O</i>	<i>NH<sub>4</sub>NO<sub>2</sub></i>
2.30143	2.29869	2.29849
29890	940	89
29816		
30182		

The respective means and estimated standard errors, the last in units of the last decimal, and the standard deviations are as follows:

*From air.*

Method 1.	2.31025 ± 3.7	$s = 8.2$
2.	2.31004 ± 6.7	$s = 13.4$
3.	2.31021 ± 5.5	$s = 9.5$

*By chemical methods.*

Method 1.	2.30008 ± 91	$s = 182$
2.	2.29904 ± 35	$s = 50$
3.	2.29869 ± 20	$s = 28$

The variation of  $s$  is striking. This is to be expected when several of the series are so short. It is plain, however, that the variability for chemical nitrogen is greater than for atmospheric nitrogen. The greatest discrepancy in the two sets is that between chemical methods 1 and 3, and can be tested by the test of 5.44; since a pair of means have been estimated we have to replace  $n_1$  by  $\nu_1 = 3$ ,  $n_2$  by  $\nu_1 = 1$ . At these values the accuracy of the approximation 5.44 (10) is of course somewhat doubtful, but we may as well see what it leads to. Here

$$e^s = 182/28 = 6.5,$$

and we find  $K = 1.9$ . As this is a selected value there seems to be no immediate need to suppose the standard error of one determination to have varied within either the set from air or the chemical set. We therefore combine the data and find the following values.

	<i>Mean</i>	<i>s</i>	<i>ν</i>	<i>s<sup>2</sup>/n</i>
From air . . . . .	2.31017 ± 0.000040	13.7	11	15.6
By chemical methods .	2.29947 ± 0.00048	137.9	7	2378.2
	0.01070			

First compare the values of  $s$ . Here  $e^2 = 10.0$ ,

$$K \doteq \left(\frac{\pi 11 \times 7}{2 \cdot 18}\right)^{1/2} \frac{100}{10.0} 10.0^7 \left(\frac{18}{7 \times 100 + 11}\right)^9 = 7.8 \times 10^{-7}.$$

The existence of a difference between the accuracies of the determinations for atmospheric and chemical nitrogen is therefore strongly confirmed. Finally, we apply 5.42 to test the difference of the means; taking the unit as 1 in the fifth decimal we get

$$\begin{aligned} P(q | \theta H) &\propto 2.1 \times 10^{-9}, & P(q_1 | \theta H) &\propto 0.12 \times 10^{-4}, \\ P(q_2 | \theta H) &\propto 1.0 \times 10^{-4}, & P(q_{12} | \theta H) &\propto 1.1 \times 10^{-4}, \\ \frac{P(q | \theta H)}{P(q_1 \vee q_2 \vee q_{12} | \theta H)} &\doteq 0.92 \times 10^{-5}. \end{aligned}$$

The existence of a systematic difference between the densities is therefore established. In this case the systematic difference is about eight times the larger standard error of one observation.

A very rough discussion can be done by the methods of contingency. The mean of all the data is 2.30978; all the 12 determinations for atmospheric nitrogen are more than this, all 8 for chemical nitrogen less. The use of a mean for comparison ensures that there will be one more and one less than the mean; hence we can allow for one parameter by deducting one from each total and testing the contingency table  $\begin{pmatrix} 7 & 0 \\ 0 & 11 \end{pmatrix}$  for proportionality of the chances. This gives by 5.14 (10)

$$K = \frac{8!}{7!0!} \frac{7!11!11!}{0!11!18!} = \frac{1}{3978},$$

which would be decisive enough for most purposes. Many problems of measurement can be reduced to contingency ones in similar ways, and the simple result is often enough. It has the advantage that it does not assume the normal law of error. It does, however, sacrifice a great deal of information if the law is true, corresponding to an increase of the standard error above what would be got by a more accurate investigation, and therefore usually (always in my experience so far) makes  $K$  too large. Thus if the rough method gives  $K < 1$  we can assert  $q'$ , but if it gives  $K > 1$  we cannot say that the observations support  $q$  without closer investigation.

According to the results the ratio of the densities is  $1.00465 \pm 0.00021$ , effectively on 7 degrees of freedom since most of the uncertainty comes from the chemical series. The 0.5, 0.1, and 0.05 points for  $t$  are at 0.71, 1.90, and 2.36. We can compare the result with what more detailed

determinations of the composition of air give. The percentages by volume of  $N_2$  and A are 78.1 and 0.93,† giving the density ratio

$$\frac{79 \times 28 + 0.93 \times 12}{79 \times 28} = 1.00505.$$

Hence 
$$t = \frac{40}{21} = 1.9,$$

which is close to the 10 per cent. point.

The outstanding problem is to understand the great difference between the standard deviations in Rayleigh's results.

**5.5. Comparison of a correlation coefficient with a suggested value.** We have seen that even in the estimation problem different ways of looking at the correlation problem suggest different ways of taking the prior probability distribution for the correlation coefficient. If we use the representation in terms of the model of 2.5 we should naturally take uniform distribution over the range permitted. If we use the rule in terms of  $J$  we have to consider whether the old parameters should be taken as  $\sigma, \tau$  or not. These parameters have the property that for any value of  $\rho$  they give the same probability distributions for  $x, y$  separately. On the other hand, they are not orthogonal to  $\rho$ . As for the testing of a simple chance the differences are not trivial, since the outside factor would vary greatly according to the suggested value of  $\rho$ , and in different ways. The difficulty is possibly connected with the question of the validity of the model and of the normal correlation law itself. In many cases where this is used it would be reasonable to regard  $x$  and  $y$  as connected in the first place by an exact linear relation, neither of them separately satisfying anything like a normal law, but subject to small disturbances which might or might not be normal. The evaluation of  $r$  in such cases is simply a test of approximate linearity of the relation between  $x$  and  $y$  and has nothing to do with normal correlation.

Tests relating to normal correlation based on  $J$  have been worked out, but suffer from a peculiarity analogous to one noticed for sampling; if the suggested value of  $\rho$  is 1 or  $-1$ , comparison of the null hypothesis with any other value of  $\rho$  makes  $J$  infinite, and the alternative hypothesis coalesces with the null hypothesis. Accordingly it seems safer to take a uniform distribution for the prior probability of  $\rho$ . We shall see that an additional restriction enters in the comparison of two correlations, similar to one that arises for comparison of samples, and

† F. A. Paneth, *Q. J. R. Met. Soc.* **63**, 1937, 433–8. Paneth states that the second figure for A is uncertain, but the uncertainty suggested would hardly affect the comparison.

that the outside factor is always of the order of the smaller of  $n_1^{1/2}$ ,  $n_2^{1/2}$ . In the first place we suppose the distribution of chance centred on  $x = y = 0$ ; the suggested value of  $\rho$  is  $\rho_0$ . Then

$$P(q \, d\sigma d\tau | H) \propto d\sigma d\tau / \sigma \tau, \quad (1)$$

$$P(q' \, d\sigma d\tau d\rho | H) \propto d\sigma d\tau d\rho / 2\sigma \tau, \quad (2)$$

the 2 entering because the possible range of  $\rho$  is from  $-1$  to  $+1$ . The likelihoods have the same form as in 3.8, and lead to

$$P(q \, d\sigma d\tau | \theta H) \propto \frac{1}{\sigma^{n+1} \tau^{n+1} (1-\rho_0^2)^{1/2n}} \exp \left[ -\frac{n}{2(1-\rho_0^2)} \left( \frac{s^2}{\sigma^2} + \frac{t^2}{\tau^2} - \frac{2\rho_0 r s t}{\sigma \tau} \right) \right] d\sigma d\tau, \quad (3)$$

$$P(q' \, d\sigma d\tau d\rho | \theta H) \propto \frac{1}{2\sigma^{n+1} \tau^{n+1} (1-\rho^2)^{1/2n}} \exp \left[ -\frac{n}{2(1-\rho^2)} \left( \frac{s^2}{\sigma^2} + \frac{t^2}{\tau^2} - \frac{2\rho r s t}{\sigma \tau} \right) \right] d\sigma d\tau d\rho. \quad (4)$$

With the substitutions 3.8 (5) we are led to

$$P(q | \theta H) \propto \int_{-\infty}^{\infty} (1-\rho_0^2)^{1/2n} (\cosh \beta - \rho_0 r)^{-n} d\beta, \quad (5)$$

$$P(q' | \theta H) \propto \frac{1}{2} \int_{-\infty}^{\infty} \int_{-1}^1 (1-\rho^2)^{1/2n} (\cosh \beta - \rho r)^{-n} d\beta d\rho. \quad (6)$$

As we only want one term in the result it is convenient to use the substitution

$$\cosh \beta - \rho r = (1-\rho r) e^u \quad (7)$$

instead of the previous one. This leads, on integration with respect to  $u$ , to

$$P(q | \theta H) \propto \frac{(1-\rho_0^2)^{1/2n}}{(1-\rho_0 r)^{n-1/2}}, \quad (8)$$

$$P(q' | \theta H) \propto \frac{1}{2} \int_{-1}^1 \frac{(1-\rho^2)^{1/2n}}{(1-\rho r)^{n-1/2}} d\rho. \quad (9)$$

Now putting

$$r = \tanh z, \quad \rho = \tanh \zeta, \quad \rho_0 = \tanh \zeta_0, \quad (10)$$

we get

$$P(q | \theta H) \propto \frac{\cosh^{n-1/2} z}{\cosh^{1/2} \zeta_0 \cosh^{n-1/2} (\zeta_0 - z)}, \quad (11)$$

$$\begin{aligned} P(q' | \theta H) &\propto \frac{1}{2} \int_{-\infty}^{\infty} \frac{\cosh^{n-1/2} z \, d\zeta}{\cosh^{5/2} \zeta \cosh^{n-1/2} (\zeta - z)} \\ &= \left( \frac{\pi}{2n-1} \right)^{1/2} \cosh^{n-3} z \end{aligned} \quad (12)$$

for large  $n$ ;  $\zeta$  has been replaced by  $z$  in the factor  $\cosh^{1/2}\zeta$ . Hence

$$K \sim \left(\frac{2n-1}{\pi}\right)^{1/2} \frac{\cosh^{1/2}z}{\cosh^{1/2}\zeta_0 \cosh^{n-1/2}(\zeta_0-z)} \quad (13)$$

$$= \left(\frac{2n-1}{\pi}\right)^{1/2} \frac{(1-\rho_0^2)^{1/2n}(1-r^2)^{1/2(n-3)}}{(1-\rho_0 r)^{n-1/2}}. \quad (14)$$

If the distribution of chance is centred on a pair of values to be determined, instead of on  $(0, 0)$ ,  $n-1$  must be substituted for  $n$ .

As an example we may take the following seismological problem. The epicentres and times of occurrence of a number of earthquakes had been determined by Bullen and me by means of a standard table of the times of travel of the  $P$  wave to different distances. Two other phases, known as  $S$  and  $SKS$ , were studied, and their mean residuals for the separate earthquakes were found.† These varied by much more than would be expected from the standard errors found for them. Such variation might arise if the focal depths of the earthquakes were not all the same, since variation of focal depth would not affect the times of all phases equally; or if any phase was multiple and there was a tendency for observers in some cases to identify the earlier, and in others the later, of two associated movements as the phase sought. In either case the result might be a correlation between the mean  $S$  and  $SKS$  residuals when  $P$  is taken as a standard. The individual values, rounded to a second, were as follows.

$S$	$SKS$	$S$	$SKS$
-8	-10	+6	+8
-5	-10	+4	+1
-3	+1	-1	0
+3	-6	+4	0
-3	+1	0	0
+3	0	-1	-1
+2	-3	-7	-2
0	+1	-8	-10
0	-4	-3	-4
+2	0		

The means are  $-0.8$  for  $S$  and  $-2.0$  for  $SKS$ . Allowing for these we find

$$\begin{aligned} \sum (x-\bar{x})^2 &= 313, & \sum (y-\bar{y})^2 &= 376, & \sum (x-\bar{x})(y-\bar{y}) &= +229; \\ s &= 4.06, & t &= 4.45, & r &= +0.667. \end{aligned}$$

There are 19 determinations and a pair of means have been eliminated.

† Jeffreys, *Bur. Centr. Intern. Séism. Assn., Trav. Sci.* **14**, 1936, 58.

Hence  $n$  in (14) must be replaced by 18. If there was no association between the residuals we should have hypothesis  $q$ , with  $\rho = 0$ ; and we find

$$K = \left(\frac{35}{\pi}\right)^{1/2} (1 - 0.667^2)^{7.5} = 0.040.$$

Thus the observations provide 25 to 1 odds on association. Further work has to try to find data that will decide between possible explanations of this association (it has appeared that both the above suggestions contain part of the truth), but for many purposes the mere fact of association is enough to indicate possible lines of progress. The later work is an instance of the separation of a disjunction as described in 1.61. Had  $K$  been found greater than 1 it would have indicated no association and both suggested explanations of the variations would have been ruled out. The tables used for comparison in obtaining the above data have been found to need substantial corrections, varying with distance and therefore from earthquake to earthquake, since the bulk of the stations observing  $S$  were at very different distances; allowance for these corrections would have made the correlation much closer. The corresponding correlations found in two later comparisons were  $+0.95$  and  $+0.97$ .†

**5.51. Comparison of correlations.** Correlations may be found from two sets of data, and the question may then arise whether the values are consistent with the true correlations being the same in both populations. We take the case where two standard errors have to be found separately for each set. On hypothesis  $q$  the true correlation is  $\rho$ , to be found from the combined data; on  $q'$  it is  $\rho_1$  in the first set and  $\rho_2$  in the second. Let the numbers of observations in the two sets be  $n_1$  and  $n_2$ , where  $n_1 > n_2$ . In accordance with the rule that the parameter  $\rho$  must appear in the statement of  $q'$ , and having regard to the standard errors of the estimates of  $\rho_1$  and  $\rho_2$ , we may define  $\rho$  on  $q'$  by

$$(n_1 + n_2)\rho = n_1\rho_1 + n_2\rho_2. \quad (1)$$

As  $\rho_2$  ranges from  $-1$  to  $+1$ , for given  $\rho$ ,  $\rho_1$  ranges from

$$\{(n_1 + n_2)\rho + n_2\}/n_1 \quad \text{to} \quad \{(n_1 + n_2)\rho - n_2\}/n_1.$$

Both are admissible values if

$$-\frac{n_1 - n_2}{n_1 + n_2} < \rho < \frac{n_1 - n_2}{n_1 + n_2}, \quad (2)$$

† *M.N.R.A.S. Geophys. Suppl.* 4, 1938, 300.

and the permitted range of  $\rho_2$  is 2. But if

$$\rho > \frac{n_1 - n_2}{n_1 + n_2}, \quad (3)$$

$$\rho_1 \text{ will be } +1 \text{ for } n_2 \rho_2 = (n_1 + n_2)\rho - n_1 \quad (4)$$

and the permitted range for  $\rho_2$  is from this value to 1, a range of  $(n_1 + n_2)(1 - |\rho|)/n_2$ . This will apply also if  $\rho$  is too small to satisfy (2). Denote the permitted range of  $\rho_2$  by  $c$ . Then the prior probabilities are

$$P(q \, d\sigma_1 d\tau_1 d\sigma_2 d\tau_2 d\rho | H) \propto d\sigma_1 d\tau_1 d\sigma_2 d\tau_2 d\rho / \sigma_1 \tau_1 \sigma_2 \tau_2, \quad (5)$$

$$P(q' \, d\sigma_1 d\tau_1 d\sigma_2 d\tau_2 d\rho d\rho_2 | H) \propto d\sigma_1 d\tau_1 d\sigma_2 d\tau_2 d\rho d\rho_2 / \sigma_1 \tau_1 \sigma_2 \tau_2 c. \quad (6)$$

The likelihoods are the products of those for the estimation problems, and we can eliminate  $\sigma_1, \tau_1, \sigma_2, \tau_2$  in terms of  $\alpha_1, \beta_1, \alpha_2, \beta_2$  as before. Then

$$P(q \, d\rho | \theta H) \propto \frac{(1 - \rho^2)^{1/2(n_1 + n_2)}}{(1 - \rho r_1)^{n_1 - 1/2} (1 - \rho r_2)^{n_2 - 1/2}} d\rho, \quad (7)$$

$$P(q' \, d\rho d\rho_2 | \theta H) \propto \frac{(1 - \rho_1^2)^{1/2 n_1} (1 - \rho_2^2)^{1/2 n_2}}{(1 - \rho_1 r_1)^{n_1 - 1/2} (1 - \rho_2 r_2)^{n_2 - 1/2}} \frac{d\rho d\rho_2}{c} \quad (8)$$

$$\propto \frac{(1 - \rho_1^2)^{1/2 n_1} (1 - \rho_2^2)^{1/2 n_2}}{(1 - \rho_1 r_1)^{n_1 - 1/2} (1 - \rho_2 r_2)^{n_2 - 1/2}} \frac{n_1 d\rho_1 d\rho_2}{(n_1 + n_2)c}, \quad (9)$$

and, using the  $\rho = \tanh \zeta$  transformation,

$$P(q | \theta H) \propto \int_{-\infty}^{\infty} \frac{\operatorname{sech} \zeta \, d\zeta}{\cosh^{n_1 - 1/2}(\zeta - z_1) \cosh^{n_2 - 1/2}(\zeta - z_2)}, \quad (10)$$

$$P(q' | \theta H) \propto \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{\operatorname{sech}^{3/2} \zeta_1 \operatorname{sech}^{3/2} \zeta_2}{\cosh^{n_1 - 1/2}(\zeta_1 - z_1) \cosh^{n_2 - 1/2}(\zeta_2 - z_2)} \frac{n_1 d\zeta_1 d\zeta_2}{(n_1 + n_2)c}. \quad (11)$$

Hence

$$\begin{aligned} P(q | \theta H) \\ \propto \left( \frac{2\pi}{n_1 + n_2 - 1} \right)^{1/2} \operatorname{sech} \frac{(n_1 - \frac{1}{2})z_1 + (n_2 - \frac{1}{2})z_2}{n_1 + n_2 - 1} \exp \left\{ - \frac{(n_1 - \frac{1}{2})(n_2 - \frac{1}{2})(z_1 - z_2)^2}{2(n_1 + n_2 - 1)} \right\}. \end{aligned} \quad (12)$$

$$P(q' | \theta H) \propto \frac{2\pi n_1}{(n_1 - \frac{1}{2})^{1/2} (n_2 - \frac{1}{2})^{1/2} (n_1 + n_2)c} \operatorname{sech}^{3/2} z_1 \operatorname{sech}^{3/2} z_2, \quad (13)$$

$$\begin{aligned} K = \left( \frac{(n_1 - \frac{1}{2})(n_2 - \frac{1}{2})}{2\pi(n_1 + n_2 - 1)} \right)^{1/2} \frac{(n_1 + n_2)c}{n_1} \operatorname{sech} \left\{ \frac{(n_1 - \frac{1}{2})z_1 + (n_2 - \frac{1}{2})z_2}{n_1 + n_2 - 1} \right\} \times \\ \times \cosh^{3/2} z_1 \cosh^{3/2} z_2 \exp \left\{ - \frac{(n_1 - \frac{1}{2})(n_2 - \frac{1}{2})(z_1 - z_2)^2}{2(n_1 + n_2 - 1)} \right\}. \end{aligned} \quad (14)$$



A little simplification is possible if we remember that a test will be needed only if  $n_1$  and  $n_2$  are both rather large, and then the critical value will be for a rather small value of  $z_1 - z_2$ . We can therefore introduce a mean value given by

$$(n_1 + n_2 - 1)z = (n_1 - \frac{1}{2})z_1 + (n_2 - \frac{1}{2})z_2; \quad (15)$$

and, nearly,  $\rho = \tanh z \quad (16)$

and  $\frac{(n_1 + n_2)c}{n_1} = \begin{cases} \frac{2(n_1 + n_2)}{n_1} & (|\rho| < \frac{n_1 - n_2}{n_1 + n_2}), \\ \frac{(n_1 + n_2)^2}{n_1 n_2} (1 - |\rho|) & (|\rho| > \frac{n_1 - n_2}{n_1 + n_2}), \end{cases} \quad (17)$

$$K = \left[ \frac{(n_1 - \frac{1}{2})(n_2 - \frac{1}{2})}{2\pi(n_1 + n_2 - 1)} \right]^{1/2} \frac{(n_1 + n_2)c}{n_1} \cosh^2 z \exp \left[ -\frac{(n_1 - \frac{1}{2})(n_2 - \frac{1}{2})(z_1 - z_2)^2}{2(n_1 + n_2 - 1)} \right]. \quad (19)$$

A further and permissible approximation will be got by identifying  $n_1$  and  $n_1 - \frac{1}{2}$ ,  $n_2$  and  $n_2 - \frac{1}{2}$  in the outside factors; we can take these as

$$\left\{ \frac{2(n_2 - \frac{1}{2})(n_1 + n_2 - 1)}{\pi(n_1 - \frac{1}{2})} \right\}^{1/2} \quad (|\rho| < \frac{n_1 - n_2}{n_1 + n_2}), \quad (20)$$

$$\left\{ \frac{(n_1 + n_2 - 1)^3}{2\pi(n_1 - \frac{1}{2})(n_2 - \frac{1}{2})} \right\}^{1/2} (1 - |\rho|) \quad (|\rho| > \frac{n_1 - n_2}{n_1 + n_2}). \quad (21)$$

The tests given above for normal correlation can be adapted immediately to rank correlation. It would be necessary only to calculate

$$1.0472(1 + 0.042\rho^2 + 0.008\rho^4 + 0.002\rho^6)$$

for the estimated  $\rho$ . Then in the outside factor of (14) we should divide by this expression, and in the exponent we should divide by its square, in accordance with the form of the approximation 5.0 (10). The correction is small enough for the effect of error in it to be regarded as negligible.

**5.6. The intraclass correlation coefficient.** This arises when we have a number of classes of  $k$  members each. If there is a component variation common to all members of a class, with standard error  $\tau$ , about some general value, and superposed on it is a variation with standard error  $\sigma'$ , the ratio of the two can be estimated from the ratio of the variation between the class means to the variation within the classes.

In the case  $k = 2$ , the expectation of the squared difference between members of the same pair is  $2\sigma'^2$ , that between members of different pairs  $2(\sigma'^2 + \tau^2) = 2\sigma^2$ . By analogy with the simple correlation coefficient we may introduce a correlation  $\rho$ , and if  $x$  and  $y$  are members of the same pair and  $E$  denotes expectations given the parameters,

$$\begin{aligned} E(x-y)^2 &= E(x^2) + E(y^2) - 2E(xy) \\ &= 2(1-\rho)\sigma^2 \end{aligned}$$

and also

$$= 2\sigma'^2.$$

Hence

$$\rho = \tau^2/\sigma^2. \quad (1)$$

The last relation provides a definition of  $\rho$  even if there are many members in each class. For if there were  $k$  in each group,  $\sigma$  and  $\tau$  retain their meaning in terms of expectations, and it would still be a valid procedure to pick out two members at random from each group, and for these the same argument will hold. Thus we can always define  $\rho$  as meaning  $\tau^2/\sigma^2$ , irrespective of the number of groups and of the number of observations per group. In terms of this definition  $\rho$  cannot be negative.

Brunt,<sup>†</sup> following Kapteyn, analyses the meaning of the correlation coefficient in general by regarding  $m$  as the number of component disturbances common to  $x$  and  $y$ , while  $n$  are independent. The correlation  $\rho$  would then be equal to  $m/(m+n)$ , and could be interpreted as a ratio capable of being estimated by sampling, with its prior probability uniformly distributed from 0 to 1. This appears to be a valid analysis of the intraclass correlation. Thus in the correlation of height between brothers it may be supposed that there is an inherited part common to both, on which random variations due to segregation are superposed. Negative values are excluded on such an analysis; to include them we need the extended analysis given in 2.5. But there seem to be many cases where this kind of analysis is valid, and there is a close analogy between the ordinary and intraclass correlation coefficients.

The conditions contemplated in the hypotheses of intraclass correlation arise in two types of case. One is illustrated by the comparison of brothers just mentioned, where members of different families may be expected to differ, on the whole, more widely than members of the same family. In agricultural tests on productivity different specimens

<sup>†</sup> *Combination of Observations*, 1931, p. 171.

are expected to differ more if they belong to different varieties than to the same variety. In these cases the comparison is a method of positive discovery, though in practice the existence of intraclass correlation is usually so well established already by examination of similar cases that the problem is practically one of estimation. In physics the problem is, perhaps, more often one of detecting unforeseen disturbances. Groups of observations made in the same way may yield independent estimates of a parameter, with uncertainties determined from their internal consistency; but when the separate estimates are compared they may differ by more than would be expected if these uncertainties are genuine. Sometimes such discrepancies lead to new discoveries; more often they only serve as a warning that the apparent accuracies are not to be trusted. Doubts are often expressed about the legitimacy of combining large numbers of observations and asserting that the uncertainty of the mean is  $n^{-1/2}$  times that of one observation. This statement is conditional on the hypothesis that the errors follow a normal law and are all independent. If they are not independent, further examination is needed before we can say what the uncertainty of the mean is. The usual physical practice is to distinguish between 'accidental' errors, which are reduced according to the usual rule when many observations are combined, and 'systematic' errors, which appear in every observation and persist in the mean. Since some systematic errors are harmonic and other variations, which are not constant, but either are predictable or may become so, an extended definition is desirable. We shall say that *a systematic error is a quantity associated with an observation, which, if its value was accurately known for one observation, would be calculable for all others.* But even with this extended meaning of 'systematic error' there are many errors that are neither accidental nor systematic in the senses stated. Personal errors of observation are often among them. It is known that two observers of star transits, for instance, will usually differ in their estimates, one systematically recording the transit earlier or later than the other. Such a difference is called the *personal equation*. If it was constant it would come within the definition of systematic error, and is usually treated as such; it is determined by comparing with a standard observer or with an automatic recording machine, and afterwards subtracted from all readings made by the observer. Karl Pearson† carried out some elaborate experiments to test whether errors of observation could be treated in this way, as a combination of a random error with a constant systematic error for

† *Phil. Trans. A*, 198, 1902, 235–99.

each observer. The conditions of the experiments were designed so as to imitate those that occur in actual astronomical observations. One type consisted of the bisection of a line by eye, the accuracy being afterwards checked by measurement. The other was essentially observation of the time of an event, the recorded time being compared with an automatic record of the event itself. The conditions resembled, respectively, those in the determination of the declination and the time of transit of a star with the transit circle. For each type of observation there were three observers, who each made about 500 observations. When the observations were taken in groups of 25 to 30 it was found that the means fluctuated, not by the amounts that would correspond to the means of 25 to 30 random errors with the general standard error indicated by the whole series, but by as much as the means of 2 to 15 independent observations should. The analysis of the variation of the observations into a constant systematic error and a random error is therefore grossly insufficient. The non-random error was not constant but reversed its sign at irregular intervals. It would resemble the kind of curve that would be obtained if numbers  $-5$  to  $+5$ , repetitions being allowed, were assigned at random at equal intervals of an argument and a polynomial found by interpolation between them. There is an element of randomness, but the mere continuity of the function implies a correlation between neighbouring interpolated values.

I shall speak of *internal correlation* as including intraclass correlation and also correlations similar to those just described.

Internal correlation habitually produces such large departures from the usual rule that the standard error of the mean is  $n^{-1/2}$  times that of one observation that the rule should never be definitely adopted until it has been checked. In a series of observations made by the same observer, and arranged in order of time, internal correlation is the normal thing, and at the present stage of knowledge hardly needs a significance test any longer. It practically reduces to a problem of estimation. The question of significance arises only when special measures have been taken to eliminate the correlation and we want to know whether they have been successful. Thus 'Student' writes:† 'After considerable experience, I have not encountered any determination which is not influenced by the date on which it is made; from this it follows that a number of determinations of the same thing made on the same day are likely to lie more closely together than if the repetitions had been made on different days. It also follows that if the

† Quoted by E. S. Pearson, *Biometrika*, 30, 1939, 228.

probable error is calculated from a number of observations made close together in point of time, much of the secular error will be left out and for general use the probable error will be too small. Where, then, the materials are sufficiently stable, it is well to run a number of determinations on the same material through any series of routine determinations which have to be made, spreading them over the whole period.' He is speaking of physical and chemical determinations. In astronomy an enormous reduction of uncertainty, by factors of 10 or 100, is achieved by combining large numbers of observations. But astronomers know by experience that they must be on the look-out for what they call systematic errors, though many of them would come under what I call internal correlation. They arrange the work so that star-positions are compared with other stars on the same plate, so that any tendency to read too high or to one side will cancel from the differences, even though it might be reversed on the next plate measured; the scale of the plate is determined separately for each plate by means of the comparison stars; special care is taken to combine observations in such a way that possible errors with daily or annual periods will not contribute systematically to the quantity to be determined; as far as possible observers are not aware what sign a systematic effect sought would have on a particular plate; and so on. In seismology many of the great advances of the past have been made by 'special studies', in which one observer collects the whole of the records of an earthquake, reads them himself, and publishes the summaries. There is here a definite risk of some personal peculiarity of the observer appearing in every observation and leading to a spurious appearance of accuracy. Bullen and I dealt with this, in the first place, by using the readings made at the stations themselves; thus any personal peculiarity would affect only one observation for each phase for each earthquake, and the resulting differences would contribute independently and could be treated as random. In the design of agricultural experiments Fisher and his followers are in the habit of eliminating some systematic ground effects as accurately as possible; the rest would not necessarily be random, but are deliberately made to contribute at random to the estimates of the effects actually sought, by randomizing the design as far as is possible consistently with the normal equations for the main effects being orthogonal.

As a specimen of the kind of results obtainable with such precautions we may take the comparisons of the times of the *P* wave in European and North American earthquakes, for distances from  $22.5^\circ$  to  $67.5^\circ$ ;

mean residuals are given against a trial table. Unit weight means a standard error of 1 sec.

$\Delta$	<i>Europe</i>		<i>N. America</i>		<i>Difference</i>	<i>Weight</i>	$\chi^2$
	<i>Mean</i>	<i>Weight</i>	<i>Mean</i>	<i>Weight</i>			
22.5	-0.2	4.7	+1.0	0.6	+0.8	0.5	0.3
23.5	-0.8	6.3	-0.1	0.6	+0.3	0.5	0.0
24.5	-1.1	3.1	+1.0	0.5	+1.7	0.4	1.2
25.5	-0.7	3.1	-0.2	0.9	+0.1	0.7	0.0
26.5	+0.3	2.7	+0.1	1.0	-0.6	0.7	0.3
27.5	-1.0	0.8	+0.3	1.2	+0.9	0.5	0.4
29.0	-0.6	4.5	+0.3	2.0	+0.5	1.4	0.4
31.5	-0.2	5.3	+0.7	2.6	+0.5	1.7	0.4
34.5	-1.8	3.1	-0.6	2.8	+0.8	1.5	1.0
37.5	-0.8	1.8	+0.8	2.1	+1.2	1.0	1.4
40.5	+0.9	1.1	-0.5	1.3	-1.8	0.6	2.0
43.5	-0.7	1.9	-1.4	0.8	-1.1	0.6	0.7
46.5	-1.2	3.0	-1.5	1.0	-0.7	0.8	0.4
49.5	-1.8	1.6	-1.4	0.8	0.0	0.5	0.0
52.5	-1.0	2.5	-2.8	1.0	-2.2	0.7	3.4
55.5	-0.7	1.9	-2.5	1.1	-2.2	0.7	3.4
58.5	-1.0	1.2	-1.4	0.3	-0.8	0.3	0.2
62.5	-1.2	1.4	-0.9	2.5	-0.1	0.9	0.1
67.5	-1.3	1.2	-0.8	3.3	+0.1	0.9	0.1
							15.7

A constant systematic difference is to be expected, corresponding to a slight difference in the way of estimating the origin times, arising from the fact that the distributions of weight outside this range are very different. The weighted mean of the difference is  $+0.4s. \pm 0.3s.$  This is added to the European mean and the result subtracted from the North American one. The results are given as 'difference', with the corresponding weights. Then

$$\chi^2 = \sum (\text{weight})(\text{difference})^2 = 15.7$$

on 19 entries, from which one parameter has been determined, so that the expectation of  $\chi^2$  is 18 on the hypothesis of randomness.

The distribution of signs at first sight suggests a systematic variation, but we notice that up to 31.5° the whole weight of the 8 differences is 6.4, and the weighted mean  $+0.45 \pm 0.40$ , which is not impressive. The last five give  $-0.91 \pm 0.58$ . The magnitude of the differences is, in fact, unusually small in the early part of the table, as we see from the fact that the largest contribution to  $\chi^2$  is 1.2. There is no contribution larger than 3.4, but on 19 entries we should have been prepared to find one greater than 4.0 on the hypothesis of randomness.

**5.61. Systematic errors: further discussion.** For simplicity we

may take the very common case where the systematic error is an additive constant. Now what can such a systematic error mean in terms of our theory? The true value, for our purposes, has been identified with the location parameter of the law of error, and the best estimate of this is definitely the mean. If, subject to it, the errors are independent, its uncertainty is correctly given by the usual formula, and we have seen how to correct it if they are not. *Systematic error has a meaning only if we understand by the true value something different from the location parameter. It is therefore an additional parameter, and requires a significance test for its assertion.* There is no epistemological difference between the Smith effect and Smith's systematic error; the difference is that Smith is pleased to find the former, while he may be annoyed at the discovery of the latter. Now with a proper understanding of induction there is no need for annoyance. It is fully recognized that laws are not final statements and that inductive inferences are not certain. The systematic error may be a source of considerable interest to his friend Smythe, an experimental psychologist. The important thing is to present the results so that they will be of the maximum use. This is done by asserting no more adjustable parameters than are supported by the data, and the best thing for Smith to do is to give his location parameter with its uncertainty as found from his observations. The number of observations should be given explicitly. It is not sufficient merely to give the standard error, because we can never guarantee absolutely that the results will never be used in a significance test, and the outside factor depends on the number of observations. Two estimates may both be  $+1.50 \pm 0.50$ , but if one is based on 10 observations with a standard error of 1.5 and the other on 90,001 with a standard error of 150, they will give respectively  $K = 0.34$  and  $K = 4.3$  in a test of whether the parameter is zero. Now this difference does not correspond to statistical practice, but it does correspond to a feeling that physicists express in some such terms as 'it is merely a statistical result and has no correspondence with physical reality'. The former result would rest on about 8 observations with positive signs, and 2 with negative, an obvious preponderance, which would give  $K = 0.49$  when tested against an even chance. The latter would rest on nearly equal numbers of observations with positive and negative signs. I think that the physicist's feeling in this is entitled to respect, and that the difference in the values of  $K$  gives it a quantitative interpretation. The mean of a large number of rough observations may have the same value and the same standard error as that of a smaller number of

accurate observations, and provided that the independence of the errors is adequately checked it is equally useful in an estimation problem; but it provides much less ground for rejecting a suggestion that the new parameter under discussion is zero when there is such a suggestion. Ultimately the reason is that the estimate is a selection from a wider range of possible values consistent with the whole variation of the observations from 0, and the difference in the values of  $K$  represents the allowance for this selection.

Now systematic differences between experiments with different methods, and even between different experimenters apparently using the same method, do exist. It is perfectly possible that what Smith does measure is something different from what he sets out to measure, and the difference is his systematic error. The quantity to be estimated may indeed be different in kind from the one actually measured. A meteorologist wants to know the atmospheric pressure, but what he observes is the height of a column of mercury. The conversion requires the use of a hydrostatic law, which is not questioned, but it involves the local value of gravity and the temperature, which enters through the density of the mercury. Allowing for the differences between these and some standard values is the removal of a calculable, and therefore a systematic, error. An astronomer wants the direction of a star, as seen from the centre of the earth; but the observed direction is affected by refraction, and the latter is calculated and allowed for. The only increase of the uncertainty involved in applying such a correction represents the uncertainty of the correction itself, which is often negligible and can in any case be found.

The problem that remains is, how should we deal with possible systematic errors that are *not* yet established and whose values are unknown? A method often adopted is to state possible limits to the systematic error and combine this with the apparent uncertainty. If the estimate is  $a \pm s$ , and a systematic error may be between  $\pm m$  (usually greater than  $s$ ), the observer may reckon the latter as corresponding to a standard error of  $m/\sqrt{3}$  and quote his uncertainty as  $\pm(s^2 + \frac{1}{3}m^2)^{1/2}$ ; or with a still more drastic treatment he may give it as  $\pm(s+m)$ . Either treatment seems to be definitely undesirable. If the existence of the error is not yet established it remains possible that it is absent, and then the original estimate is right. If it exists, the evidence for its existence will involve an estimate of its actual amount, and then it should be allowed for; and the uncertainty of the corrected estimate will be the resultant of  $s$  and the determined uncertainty of



the systematic correction. In either case  $s$  has a useful function to serve, and should be stated separately and not confused with  $m$ . The possible usefulness of  $m$ , where the existence of the error is not established and its actual amount therefore unknown, is that it suggests a possible range of values for a new parameter, which may be useful in comparison with other series of observations when material becomes available to test the presence of a systematic difference. But inspection of our general approximate formula shows that the statement of  $m$  will go into the outside factor, not into the standard error. If the standard error is inflated by  $m$  the result will be to increase the uncertainty unjustifiably if the suggested difference is not revealed by the accurate test; and to fail to reveal a difference at all when the test should show it and lead to an estimate of its amount. In either case the inclusion of  $m$  in the uncertainty leads to the sacrifice of information contained in the observations that would be necessary to further progress (cf. 5.63). A separate statement of the possible range of the systematic error may be useful if there is any way of arriving at one, but it must be a separate statement and not used to increase the uncertainty provided by the consistency of the observations themselves, which has a value for the future in any case. In induction there is no harm in being occasionally wrong; it is inevitable that we shall be. But there is harm in stating results in such a form that they do not represent the evidence available at the time when they are stated, or make it impossible for future workers to make the best use of that evidence.

**5.62. Estimation of intraclass correlation.** In most treatments of this problem, including the one in the first edition of this book, the classes compared have been supposed equal in number. In such cases  $K$  can be reduced to a single integral. This condition is satisfied in balanced designs, such as are often used in biological experiments. In other applications it is rarely satisfied. However carefully an astronomer designs his observing programme it will generally be interrupted by cloud. Even in the comparison of brothers there is no theoretical reason for taking the same number from every family; the reason is only to make the analysis fairly easy. But it is usual for the scatter within the groups to give an estimate of the random error sufficiently accurate to be taken as a definite determination of  $\sigma$ . We suppose then that there is a general location parameter  $\lambda$ ; that there are  $m$  groups of observations, the number in the  $r$ th group being  $k_r$ , and that there is a location parameter  $\lambda_r$  associated with the group whose probability distribution about  $\lambda$  is normal with standard error  $\tau$ ; and that within each group

the observed values are random with standard error  $\sigma$  about  $\lambda_r$ . The uncertainty of  $\sigma$  is taken as negligible. We suppose the separate values  $\lambda_r - \lambda$ , given  $\tau$ , to be independent. This is the fundamental distinction between intraclass correlation and systematic variation. The data are the group means  $x_r$ . According to the hypotheses

$$x_r = \lambda \pm \sqrt{(\tau^2 + \sigma^2/k_r)} \quad (1)$$

and the likelihood is

$$L = (2\pi)^{-1/2m} \prod_r (\tau^2 + \sigma^2/k_r)^{-1/2} \exp\left\{-\frac{1}{2} \sum_r \frac{(x_r - \lambda)^2}{\tau^2 + \sigma^2/k_r}\right\} \prod dx_r. \quad (2)$$

Then we have to estimate  $\lambda$  and  $\tau$ . We have

$$\frac{\partial}{\partial \lambda} \log L = \sum \frac{k_r(x_r - \lambda)}{\sigma^2 + k_r \tau^2}, \quad (3)$$

$$\frac{\partial}{\partial \tau^2} \log L = -\frac{1}{2} \sum \frac{k_r}{\sigma^2 + k_r \tau^2} + \frac{1}{2} \sum \frac{k_r^2(x_r - \lambda)^2}{(\sigma^2 + k_r \tau^2)^2}. \quad (4)$$

Putting these zero we have the maximum likelihood equations for  $\lambda$  and  $\tau^2$ . To get the uncertainties we need also the second derivatives

$$\frac{\partial^2}{\partial \lambda^2} \log L = -\sum \frac{k_r}{\sigma^2 + k_r \tau^2}, \quad (5)$$

$$\frac{\partial^2}{(\partial \tau^2)^2} \log L = \frac{1}{2} \sum \frac{k_r^2}{(\sigma^2 + k_r \tau^2)^2} - \sum \frac{k_r^3(x_r - \lambda)^2}{(\sigma^2 + k_r \tau^2)^3}. \quad (6)$$

The posterior probability distribution of  $\lambda$  will not reduce to a simple  $t$  rule. If  $\tau$  was 0 it would be normal with standard error  $(\sum k_r)^{-1/2} \sigma$ . If  $\sigma$  was 0 it would follow a  $t$  rule with  $m-1$  degrees of freedom. We are concerned with intermediate cases, and may expect that the distribution will resemble a  $t$  rule with more than  $m-1$  degrees of freedom. To estimate the number we form the corresponding derivatives for the normal law. Here we have

$$\log L = -n \log \sigma - \frac{n}{2\sigma^2} \{(\bar{x} - \lambda)^2 + s'^2\}, \quad (7)$$

$$\frac{\partial}{\partial \lambda} \log L = \frac{n}{\sigma^2} (\bar{x} - \lambda), \quad (8)$$

$$\frac{\partial}{\partial \sigma^2} \log L = -\frac{n}{2\sigma^2} + \frac{n}{2\sigma^4} \{s'^2 + (\bar{x} - \lambda)^2\}, \quad (9)$$

$$\frac{\partial^2}{\partial \lambda^2} \log L = -\frac{n}{\sigma^2}, \quad (10)$$

$$\frac{\partial^2}{(\partial \sigma^2)^2} \log L = \frac{n}{2\sigma^4} - \frac{n}{\sigma^6} \{s'^2 + (\bar{x} - \lambda)^2\}. \quad (11)$$

(8) and (9) vanish when  $\lambda = \bar{x}$ ,  $\sigma = s'$ ; and then (10) becomes  $-n/s'^2$  and (11) becomes  $-n/2s'^4$ . Hence, to the second order in departures from the maximum likelihood solution,

$$\log L = \text{constant} - \frac{n}{2s'^2}(\bar{x} - \lambda)^2 - \frac{n}{4s'^4}(\sigma^2 - s'^2)^2.$$

But it is simply the uncertainty of  $\sigma$  that produces the departure of the  $t$  rule from the normal. Consider then the value  $-A$  taken by (10) when  $\sigma^2 = s'^2$ , and the value  $-B$  taken when

$$\sigma^2 = s'^2 \left( 1 + \sqrt{\frac{2}{n}} \right).$$

We have 
$$\frac{A}{B} = 1 + \sqrt{\left(\frac{2}{n}\right)}, \quad n = \frac{2}{(A/B - 1)^2}.$$

Then the number of degrees of freedom is  $n-1$ ; and the  $s_a$  of the  $t$  rule is given by

$$s_a^2 = \frac{s'^2}{n-1} = \frac{n}{(n-1)A}.$$

This can be immediately adapted to (5) and (6). We work out (6) for the maximum likelihood solution. (5) for this solution is  $-A$ ; (5) with  $\tau^2$  increased by its standard error indicated by (6) is  $-B$ . An approximate  $t$  rule for  $\lambda$  follows.

The following data on the correction to the constant of nutation, derived from a combination of data by Sir H. Spencer Jones,<sup>†</sup> provide an illustration. The separate equations of condition are from comparisons of different pairs of stars. The unit is taken as 0.01"; the standard error for unit weight derived from internal comparisons is 7.7. The weights have been rounded to the nearest unit.

$k_r$	$x_r$	$k_r(x_r - \bar{x}_r)^2$
44	-2.02	325
25	+3.52	200
23	+4.17	293
25	+0.11	8
8	-1.73	47
5	+4.89	90
3	+4.28	39
18	-0.82	41
		1043

The weighted mean is +0.69 and gives  $\chi^2 = 1043/7.7^2 = 16.9$  on 7 degrees of freedom. This is beyond the 2 per cent. point, and is enough to arouse suspicion. The original series, before they were combined to

<sup>†</sup> *M.N.R.A.S.* 98, 1938, 440-7.

give the above estimates, had shown similar discrepancies, one of them being beyond the 0.1 per cent. point. There is further confirmation from the distribution of the contributions to  $\chi^2$ . For random variation these should not be correlated with  $k_r$ . Actually the three largest contributions come from three of the four largest  $k_r$ , which is what we should expect if intraclass correlation is present. We therefore proceed to estimate  $\tau^2$ .

To get an upper estimate we treat all the values as of equal weight, thus neglecting  $\sigma^2$ . The simple mean is  $+1.55$ —which is a warning that if  $\tau$  is not taken into account there may be a serious error in the estimation of  $\lambda$ —and the residuals give  $\tau^2 = 8.8$ . This is too high since the variation includes the part due to  $\sigma$ .

We write  $w_r = k_r/(\sigma^2 + k_r\tau^2)$ .

For purposes of computation  $\lambda$  is taken as  $\lambda_0 = +1.13$  (suggested by the first trial value  $\tau^2 = 6.0$ ), and  $w_r$  is worked out for several trial values of  $\tau^2$ . Results are as follows.

$\tau^2$	$\sum w_r$	$\sum w(x_r - \lambda_0)$	$\sum w^2$	$\sum w^2(x_r - \lambda_0)^2$	$\sum w^2(x_r - \lambda_0)^2$	$\lambda - \lambda_0$	$\sum w_r - \sum w_r^2(x_r - \lambda_0)^2$
3.0	1.151	-0.111	0.1963	1.316	0.246	-0.10	-0.165
3.5	1.060	-0.090	0.1642	1.103	0.186	-0.085	-0.043
4.0	0.984	-0.064	0.1398	0.936	0.144	-0.065	+0.048
5.0	0.865	-0.024	0.1059	0.712	—	-0.03	+0.153

By (4) we have to interpolate so that  $\sum w_r - \sum w_r^2(x_r - \lambda)^2 = 0$ . We can neglect the difference between  $\lambda$  and  $\lambda_0$ . Interpolation gives

$$\tau^2 = 3.71,$$

and the interpolated value of  $\lambda - \lambda_0$  is  $-0.075$ , hence

$$\lambda = +1.055.$$

Also

$$-\frac{\partial^2}{\partial \lambda^2} \log L = \sum w_r,$$

$$-\frac{\partial^2}{(\partial \tau^2)^2} \log L = -\frac{1}{2} \sum w_r^2 + \sum w_r^3(x_r - \lambda)^2 = +0.091.$$

Then we can take  $\tau^2 = 3.71 \pm 3.32$ . Substitute in  $\sum w_r$  for  $\tau^2 = 3.71$  and 6.0; we get respectively  $+1.02$  and  $0.77$ . Extrapolating to  $\tau^2 = 7.03$  we have  $\sum w_r = 0.66$ ,

$$n = \frac{2}{(1.02/0.66 - 1)^2} \div 7, \quad s_\lambda^2 = \frac{7}{6 \times 1.02} = 1.14.$$

Changing the unit to 1" we have the solution

$$\lambda = +0.0105'' \pm 0.0107'', \quad 6 \text{ d.f.},$$

$$\tau = 0.0193'' \pm 0.0073''.$$

This solution is given only as an illustration of the method. A discussion using expectations gave similar conclusions,<sup>†</sup> but led Spencer Jones to go more into detail. He discovered a systematic effect that had been overlooked, and on allowing for it he obtained a satisfactory agreement with the hypothesis of independence of the errors, and consequently a substantial increase in accuracy.<sup>‡</sup> His result was

$$\lambda = +0.0034'' \pm 0.0062''.$$

The question of a significance test for  $\tau$  will arise in such problems. We notice that on the hypothesis  $\tau = 0$  a mean  $x_r$  has a standard error  $\sigma/\sqrt{k_r}$ , and for other  $\tau$  one of  $(\sigma^2/k_r + \tau^2)^{1/2}$ . Hence, for small  $\tau^2$ ,  $J$  will be of the order of magnitude of  $\tau^2$ , not  $\tau$ . In applying the approximate form for  $K$  we should therefore take

$$K \doteq \left(\frac{\pi n}{8}\right)^{1/2} \exp\{-\frac{1}{2}\tau^2/s_r^2\},$$

as suggested by 5.31 (5); a factor  $\frac{1}{2}$  is needed because  $\tau^2$  cannot be negative.

The determination of the constant of gravitation provides an illustration of the danger of drastic rejection of observations and of the method of combining estimates when the variation is not wholly random. C. V. Boys gave the value  $6.658 \times 10^{-8}$  c.g.s. But P. R. Heyl,<sup>§</sup> quoting Boys's separate values, points out that the simple mean, apart from the factor  $10^{-8}$ , is 6.663. There were nine determinations, of which all but two were rejected, so that the final result was the mean of only two observations with an unknown standard error. Even if these had been the only two observations the uncertainty of the standard error would have a pronounced effect on the posterior probability distribution; but when they are selected out of nine the accuracy is practically impossible to assess. Heyl made three sets of determinations, using balls of gold, platinum, and optical glass respectively. The summaries are as follows, with the estimated standard errors.

						<i>n</i>
Boys	.	.	.	.	$6.663 \pm 0.0023$	9
Heyl						
Gold	.	.	.	.	$6.678 \pm 0.0016$	6
Platinum	.	.	.	.	$6.664 \pm 0.0013$	5
Glass	.	.	.	.	$6.674 \pm 0.0027$	5

The estimates are plainly discrepant. Heyl has tested the possibility of a real difference between the constant of gravitation for different

<sup>†</sup> *M.N.R.A.S.* **99**, 1939, 206-10.

<sup>‡</sup> *Ibid.*, pp. 211-16.

<sup>§</sup> *Bur. Standards Res. J.* **5**, 1930, 1243-90.

substances by means of the Eötvös balance and finds none; and there is no apparent explanation of the differences. They are so large that we may compute the simple mean at once; it is 6.670, and the sum of squares of the residuals is  $165 \times 10^{-6}$ , of which the known uncertainties account for  $17 \times 10^{-6}$ . The standard error of an entire series can then be taken as  $(148/3)^{1/2} \times 10^{-3} = 0.0070$ . Combining this with the known uncertainties we get for the respective  $\sigma^2$ :  $10^{-8}(54, 52, 51, 56)$ . An improved value could be got by computing a revised mean with the reciprocals of these as weights, but they are so nearly equal that the simple mean will be reproduced. The standard error can then be taken as

$$10^{-3} \left( \frac{165}{3 \times 4} \right)^{1/2} = 3.7 \times 10^{-3},$$

and the result is  $10^{-8}(6.670 \pm 0.0037)$ . The result is, however, virtually based on only three degrees of freedom; the root-mean-square estimate of uncertainty would be

$$10^{-3} \left( \frac{165}{1 \times 4} \right)^{1/2} = 6.4 \times 10^{-3},$$

and this would be the safest to use in matters where the chief uncertainty arises from the constant of gravitation.

**5.63. Suspiciously close agreement.** The tendency of either internal correlation or of a neglected systematic effect is in general to increase  $\chi^2$  or  $z$ , and it is chiefly to this fact that these functions owe their importance. If they agree reasonably with their expectations the null hypothesis can usually be accepted without further ado. But it sometimes happens that  $\chi^2$  is much less than its expectation; an analogous result would be strongly negative  $z$  when the variation suspected of containing a systematic part is compared with the estimate of error; another is when the standard error of a series of measures is much less than known sources of uncertainty suggest. Strong opinions are expressed on this sort of agreement. Thus Yule and Kendall remark:†

‘Nor do only small values of  $P$  (the probability of getting a larger  $\chi^2$  by accident) lead us to suspect our hypothesis or our sampling technique. A value of  $P$  very near to unity may also do so. This rather surprising result arises in this way: a large value of  $P$  normally corresponds to a small value of  $\chi^2$ , that is to say a very close agreement between theory and fact. Now such agreements are rare—almost as rare as great divergences. We are just as unlikely to get very good correspondence between fact and theory as we are to get very bad correspondence and, for precisely the same reasons, we must suspect our sampling technique if we do. In short, very close agreement is *too good to be true*.

† *Introduction to the Theory of Statistics*, p. 423.

'The student who feels some hesitation about this statement may like to reassure himself with the following example. An investigator says that he threw a die 600 times and got exactly 100 of each number from 1 to 6. This is the theoretical expectation,  $\chi^2 = 0$  and  $P = 1$ , but should we believe him? We might, if we knew him very well, but we should probably regard him as somewhat lucky, which is only another way of saying that he has brought off a very improbable event.'<sup>†</sup>

Similarly, Fisher writes:<sup>‡</sup>

'If  $P$  is between 0.1 and 0.9 there is certainly no need to suspect the hypothesis tested. . . .'

'The term Goodness of Fit has caused some to fall into the fallacy of believing that the higher the value of  $P$  the more satisfactorily is the hypothesis verified. Values over 0.999 have been reported, which, if the hypothesis were true, would only occur once in a thousand trials. Generally such cases are demonstrably due to the use of inaccurate formulae, but occasionally small values of  $\chi^2$  beyond the expected range do occur, . . . In these cases the hypothesis is as definitely disproved as if  $P$  had been 0.001.'

A striking case is given by Fisher§ himself in a discussion of the data in Mendel's classical papers on inheritance. In every case the data agreed with the theoretical ratios within less than the standard errors; taking the whole together,  $\chi^2$  was 41.6 on 84 degrees of freedom, and the chance of a smaller value arising accidentally is 0.00007. The test originated in two cases where Mendel had distinguished the pure and heterozygous dominants by self-fertilization, growing ten of the next generation from each. Since the chance of a self-fertilized heterozygote giving a dominant is  $\frac{3}{4}$ , the chance that all ten would be dominants is  $(0.75)^{10} = 0.05$ , so that about 5 per cent. of the heterozygous ones would fail to be detected, and the numbers would be underestimated. Correcting for this, Fisher found that Mendel's observed numbers agreed too closely with the uncorrected ratio of one pure to two mixed dominants, while they showed a serious discrepancy from the corrected ratio. Fisher suggests that an enthusiastic assistant, knowing only too well what Mendel expected, made the numbers agree with his expectations more closely than they need, even in a case where Mendel had overlooked a complication that would lead the theoretical ratio to differ appreciably from the simple 1:2.

When there is only one degree of freedom to be tested a very close agreement is not remarkable—if two sets of measures refer to the same thing, agreement between the estimates within the rounding-off error

<sup>†</sup> To go to the other extreme, if a man reports that he obtained a complete hand of one suit at bridge we do not believe that he did so by a random deal. It is more likely either that he is lying or that something was wrong with the shuffling.

<sup>‡</sup> *Statistical Methods*, 1936, p. 84.

<sup>§</sup> *Annals of Science* 1, 1936, 115-37.

is the most probable result, even though its probability is of the order of the ratio of the rounding-off error to the standard error of the difference. It is only when such agreements are found persistently that there is ground for suspicion. The probable values of  $\chi^2$  from 84 degrees of freedom are  $84 \pm 13$ , not 0. If the only variations from the null hypothesis were of the types we have discussed here, too small a  $\chi^2$  would always be evidence against them. Unfortunately there is another type. By some tendency to naïve notions of causality, apparent discrepancies from theory are readily reduced in the presentation of the data. People not trained in statistical methods tend to underestimate the departures that can occur by chance, a purely random result is in consequence often accepted as systematic when no significance test would accept it as such, and 'effects' make transitory appearances in the scientific journals until other workers repeat the experiments or estimate the uncertainty properly. Similarly, when the investigator believes in a theory he is predisposed to think that if a set of observations differs appreciably from expectation there is something wrong with the observations, even though a closer examination would show that the difference is no larger than would often occur by chance; and the consequence is that observations may be rejected or illegitimately modified before presentation. This tendency is the more dangerous because it may be completely unconscious. In Mendel's experiments, where there were theoretical ratios to serve as a standard, the result would be too small a  $\chi^2$ , which is what Fisher found.

A significance test for such cases on the lines of the present chapter has not been constructed. It would be most useful if the prior probability took account of previous information on human mendacity, but this has not, I think, been collected in a useful form!

5.64. Sir Arthur Eddington has claimed to have deduced theoretical values of many measurable physical quantities from purely epistemological considerations. I consider that this is at least partly because he has incorporated a great deal of observational material into what he calls epistemology;† but that is not the chief reason why the great majority of physicists hesitate to accept his arguments. At any rate it is interesting to compare the values deduced theoretically in his *Fundamental Theory* with observation. He takes the velocity of light, the Rydberg constant, and the Faraday constant as fundamental and calculates the rest from them. I give his comparisons as they stand except for powers of 10, which are irrelevant for the present purpose;

† *Phil. Mag.* (7), 32, 1941, 177–205.



uncertainties are given as 'probable errors' and the factor  $(0.6745)^2$  must be applied at some stage in the computation of  $\chi^2$ . Probable errors are given for the last figure in the observed value.

	<i>Obs.</i>	<i>P.E.</i>	<i>Calc.</i>	<i>O. - C.</i>	$(0.6745)^{-2}\chi^2$
$e/m_0c$ (deflexion)	1.75959	24	1.75953	+6	0.1
$e/m_0c$ (spectroscopic)	1.75934	28	1.75953	-19	0.5
$hc/2\pi e^2$	137.009	16	137.000	+9	0.3
$m_p/m_e$	1836.27	56	1836.34	-7	0.0
$M$	1.67339	31	1.67368	-29	0.9
$m_s$	9.1066	22	9.1092	-26	1.4
$e'$	4.8025	10	4.8033	-8	0.6
$\hbar'$	6.6242	24	6.6250	-8	0.1
$\hbar/e'$	1.3800	5	1.3797	+3	0.4
$\kappa$	6.670	5	6.6665	+3.5	0.5
$n' - H'$	0.00082	3	0.0008236	-0.4	0.0
$2H' - D'$	0.001539	2	0.0015404	-1.4	0.5
$4H - He$	0.02866	?	$0.02862 \pm 4$	+4	$\leq 1.0$
$\mathfrak{M}$	2.7896	8	2.7899	-3	0.1
$\mathfrak{M}_n$	1.935	20	1.9371	-2.1	0.0
					$\leq 6.4$

I have omitted some of Eddington's comparisons but retained, I think, all where the observed values rest on independent experiments. The result is that  $\chi^2$  is not more than 2.9, on 15 d.f. This is preposterous; the 99 per cent. point is at  $\chi^2 = 5.2$ .

It might theoretically be better not to take three constants as definitely known, but to make a least-squares solution from 18 data, taking these as unknown, using their experimental uncertainties. This would not make much difference since they are among those whose uncertainties are smallest compared with the adopted values; the only difference would be that  $\chi^2$  would be slightly reduced, remaining on 15 d.f.

Many of the observed values are based on very few degrees of freedom;  $\kappa$ , the constant of gravitation, for instance, is on 3 d.f. In these conditions the use of  $\chi^2$  as if the errors were normally distributed is seriously wrong (cf. 2.82); but the tendency of the allowance for small numbers of degrees of freedom would be to increase the expectation of  $\chi^2$ , and a more accurate test would give a larger predicted  $\chi^2$ . Thus correction of either of the obvious statistical blemishes would increase the discrepancy; and the observations agree with Eddington's theory far better than they have any business to do if that theory is right.

There are two possible explanations. The one that would occur to many physicists is that Eddington's theory is artificial throughout, and that by skilful juggling with numbers he has produced a forced

agreement. This may be so, though I should not say that his theory is at any point more artificial or less intelligible than any other statement of quantum theory. All need a complete restatement of their relations to experience, including a statement of what features in experience demand the kind of analysis that has been adopted.

The other concerns the 'probable errors' of the observed values. Many of these are not based on a statistical discussion, but include an allowance for possible systematic errors, of the kind that is deprecated in 5.61. It is quite possible that the probable errors given are systematically two or three times what a proper statistical discussion would give. In particular, some of the estimates are the results of combining several different determinations, alleged to be discrepant, but as the number of degrees of freedom of the separate determinations is never given, it is impossible to form a judgement on the existence of these discrepancies without working through the whole of the original data afresh. If the uncertainties had not been artificially inflated it is possible that a normal  $\chi^2$  would have been found. At any rate the first suggested explanation cannot be accepted until the second is excluded by a rediscussion of the experimental data.

5.65. In counting experiments the standard error is fixed by the numbers of the counts alone, subject to the condition of independence. In measurement the matter is more complicated, since observers like their standard error to be small, and it is one of the unknowns of the problem and has to be judged only from the amounts of the residuals. But actually the standard error of one observation is not often of much further interest in estimation problems; what matters most, both in estimation problems and in any significance test that may supervene, is the standard error of the estimates. Now it is easy in some types of investigation for an apparent reduction of the standard error of one observation to be associated with no reduction at all in the accuracy of the estimates. This can be illustrated by the following example. A set of dice were thrown, sixes being rejected, and 3 was subtracted from each result. Thus a set of numbers  $-2$  to  $+2$ , arranged at random, was obtained (series *A*). Differences to order 4 were found, and two smoothed sets of values *B* and *C* were obtained, one by adding  $\frac{1}{4}$  of the second difference, one by subtracting  $\frac{1}{12}$  of the fourth difference. The unsmoothed and the two smoothed series are shown below. The respective sums of the 44 squares, excluding for the series *A* the two unsmoothed values at each end, are 88,† 18.9, and 29.7. The smoothing

† This agrees exactly with expectation!

has produced a great reduction in the general magnitude of the residuals; judging by this alone the standard errors have been multiplied by 0.46 and 0.58 by the two methods. But actually, if we want a summary based on the means of more than about 5 consecutive values we have gained no accuracy at all. For if a group of successive entries in column *A* are  $x_{-2}, x_{-1}, x_0, x_1, x_2$ , method *B* will make  $x_0$  contribute

<i>A</i>	<i>B</i>	<i>C</i>	<i>A</i>	<i>B</i>	<i>C</i>	<i>A</i>	<i>B</i>	<i>C</i>
0			+2	+1.0	+0.8	-2	-1.0	-1.8
+2			0	+1.0	+1.5	-2	-1.0	-1.7
-2	-0.8	-0.8	+2	+0.5	+0.2	0	-0.5	-0.4
-1	-0.5	-0.5	-2	-0.2	-0.8	0	-0.2	-0.2
+2	+0.2	+0.3	-1	-0.5	-0.7	-1	-0.2	-0.1
-2	-1.0	-1.1	+2	+0.8	+0.8	+1	0.0	-0.2
-2	-1.0	-1.2	0	+0.8	+1.2	-1	0.0	+0.2
+2	-0.5	+0.7	+1	0.0	-0.4	+1	+0.2	+0.1
0	0.0	+0.2	-2	-0.5	-0.4	0	+0.2	+0.5
-2	-1.0	-1.2	+1	+0.2	0.0	0	-0.2	-0.2
0	-0.2	-0.2	+1	+1.0	+1.5	-1	-1.0	-0.8
+1	+0.2	+0.5	+1	+0.2	+0.2	-2	-1.0	-0.8
-1	-0.8	-0.9	-2	-1.0	-1.2	+1	-0.5	-0.8
-2	-1.0	-1.1	-1	-0.5	-0.6	-2	-0.8	-0.3
+1	0.0	-0.2	+2	+0.8	+1.0	0		
0	+0.8	+1.2	0	0.0	+0.2	-2		

$\frac{1}{4}x_0$  to the second and fourth entries and  $\frac{1}{2}x_0$  to the third; the contribution from  $x_0$  to the sum of the five remains  $x_0$ . Method *C* will make  $x_0$  contribute  $-\frac{1}{12}x_0$  to the first and fifth entries,  $\frac{1}{3}x_0$  to the second and fourth, and  $\frac{1}{2}x_0$  to the third. Again there is no change in the sum of the five. There is a little gain through the contributions from the entries for adjacent ranges, but the longer the ranges are the smaller this will be.

Now it might well happen that we have a series of observations of what should be a linear function of an independent variable, and that the above set of values *A* are the errors rounded to a unit.† The least-squares solution based on the hypothesis of the independence of the errors will be valid. If a smoothing process changes the errors to *B* or *C* the solution will be the same; but if the errors are still supposed independent the apparent accuracy will be much too high, because we know that the correct uncertainty is given by *A*. What the smoothing as in *B* does, if the error at one value is  $x_0$ , independent of adjacent values, is to make component errors  $\frac{1}{4}x_0, \frac{1}{2}x_0, \frac{1}{4}x_0$  at adjacent values. Thus, though the smoothing somewhat improves the individual values, it does so by introducing a correlation between consecutive errors; and if

† The process actually used gets them from a rectangular and not a normal distribution of chance, but this is irrelevant here.

the errors are given by  $B$  or  $C$  this departure from independence of the errors is responsible for a diminished real accuracy in comparison with the apparent accuracy obtained on the hypothesis of independence.

Now at the best the hypothesis of independence of the errors needs a check when suitable information becomes available; it is never certain. But it does often survive a test, and the estimate of uncertainty is then valid. If there is any possibility that it is true, that possibility should not be sacrificed. There is a real danger in some types of observation that spurious accuracy may be obtained by introducing a correlation between neighbouring errors. In seismological work, for instance, a careful observer may read his records again and again to make 'sure', working out his residuals after each set of readings; and in these conditions it is practically impossible for him to avoid letting his readings on one record be influenced by those at neighbouring distances. There is a further danger of accidental close agreement in the results for a few separate series; knowledge of the standard error of each series based on the hypothesis of independence prevents too high an accuracy from being asserted in such cases.

In some cases a lack of independence arising in this way can be detected by comparing determinations from different series of observations; too large a  $\chi^2$  may be found, and then the differences between the series provide a valid estimate of uncertainty, though based on fewer degrees of freedom than might have been available in the first place. But even here it may happen that previous results are used to reject observations, and then even this independence fails. If the possibility of this check is to be preserved, every series must be reduced independently. Otherwise a mistake made at the outset may never be found out.

**5.7. Test of the normal law of error.** Actual distributions of errors of observation usually follow the normal law sufficiently closely to make departures from it hard to detect with fewer than about 500 observations. Unfortunately this does not show that the treatment appropriate to the normal law is appropriate also to the actual law; the same is true for a binomial law with only three or four components, or for a triangular law, and for these the extreme observations have an importance in estimation that far exceeds any they can have on the normal law. (The binomial would of course have to be compared with a normal law with the chances grouped at equal intervals.) Many series of observations have been published as supporting the normal law.

Pearson showed in his original  $\chi^2$  paper that some of these showed such departures from the normal law as would warrant its rejection. I have myself analysed nine series for this purpose.† Six of these are from a paper by Pearson, which has already been mentioned (p. 270). W. N. Bond made a series of about 1,000 readings of the position of an illuminated slit, viewed with a travelling microscope slightly out of focus. The slit was kept fixed, but the microscope was moved well outside the range of vision after each reading, so that the errors would be as far as possible independent. The conditions resemble the measurement of a spectrum line or, apart from the shape of the object, that of a star image on a photographic plate. Later Dr. H. R. Hulme provided me with two long series of residuals obtained in the analysis of the variation of latitude observations at Greenwich. These have the special interest that they are based on observations really intended to measure something and not simply to test the normal law; but Pearson's were primarily designed to test the hypothesis that the error of observation could be regarded as the sum of a constant personal error and a random error, the test of the normal law being a secondary feature. So many lists of residuals exist that could be compared with the normal law that published comparisons are under some suspicion of having been selected on account of specially good agreement with it.

In comparison with the normal law, Type VII gives  $J$  infinite for  $m = 1$ ; Type II gives  $J$  infinite for any  $m$ , but we can modify the definition by omitting the intervals where the probability according to Type II is zero, and then  $J$  remains finite, tending to infinity only as  $m \rightarrow 1$ . It is sufficient for our purposes to use the approximate formula of 5.31. The maximum likelihood solutions for  $\mu$ , which is  $1/m$  for Type VII and  $-1/m$  for Type II, are as follows.

		$n$	$\mu$	$K$
Pearson: Bisection . . .	1	500	$+0.111 \pm 0.037$	0.31
	2	500	$+0.04 \pm 0.04$	17
	3	500	$-0.225 \pm 0.057$	0.0116
Bright line . . .	1	519	$+0.230 \pm 0.057$	0.0083
	2	519	$+0.163 \pm 0.050$	0.140
	3	519	$-0.080 \pm 0.049$	7.5
Bond . . . . .		1026	$+0.123 \pm 0.051$	2.2
Greenwich . . . . .	1	4540	$+0.369 \pm 0.020$	$10^{-78}$
	2	5014	$+0.443 \pm 0.018$	$10^{-130}$

Six of the nine series give  $K$  less than 1, three less than 0.01. Allowance

† *Phil. Trans. A*, 237, 1938, 231-71; *M.N.R.A.S.* 99, 1939, 703-9.

for selection as in 5.04 does not alter this, but the larger values of  $K$  are, of course, reduced. But there is another check. If the errors, apart from a constant personal error, were random and followed the normal law, the means of groups of 25 consecutive observations should be derived from a normal law, with standard error  $\frac{1}{5}$  of that of the whole series. If  $\gamma^2$  is the square of the observed ratio, it should be about 0.04. In every case the actual value in Pearson's series was higher; it actually ranged from 0.066 to 0.550. The test for comparison of two standard errors, with  $n_1 = 20$ ,  $n_2 = 480$ , will obviously give  $K$  much less than 1 in every case. One apparently possible explanation would be that if errors follow a Type VII law, even if they are independent, means of a finite number of observations will fluctuate more than on the normal law. If this was the right explanation  $\gamma$  should increase with  $\mu$ . The actual variation is in the other direction. Taking the values in order of decreasing  $\mu$  we have the following table.

		$\mu$	$\gamma^2$	$r$
Bright line . . .	1	+0.230	0.066	0.16
	2	+0.163	0.100	0.24
Bisection . . .	1	+0.115	0.093	0.23
	2	+0.04	0.36	0.57
Bright line . . .	3	-0.080	0.140	0.32
Bisection . . .	3	-0.225	0.550	0.72

$r$  is defined as  $\sqrt{(\gamma^2 - 0.04)}$  and is an estimate of the fraction of the standard error that persists through 25 observations. There is a correlation of -0.92 between  $\mu$  and  $r$ , which might represent a practically perfect correlation since both  $\mu$  and  $r$  have appreciable uncertainties. If we fit a linear form by least squares, treating all determinations as of equal weight, we get

$$\mu = +0.273 \pm 0.093 - (0.62 \pm 0.22)r.$$

The suggestion of these results is therefore that reduction in  $\mu$  is strongly associated with increase in the correlation between consecutive errors, and that a set of really independent errors, if there is such a thing, would satisfy a Type VII law with  $m$  probably between 2.7 and 5.5.

Bond's data would suggest limits for  $m$ , corresponding to the standard error, of 5.7 to 14; the two Greenwich series of 2.6 to 2.9 and 2.2 to 2.4. There appear to be real differences in the values of  $m$ , but this has an obvious explanation. Pearson's and Bond's series were each made by

a single observer in conditions designed to be as uniform as possible. The Greenwich observations were made by several different observers in different conditions of observation. This would naturally lead to a variation of accuracy. But if several homogeneous series of different accuracy, even if derived from the normal law, were combined and the result analysed, we should get a positive  $\mu$ . The values found from the Greenwich observations are therefore likely to be too high for uniform observing conditions. It seems that for uniform conditions, if independence of the errors can be attained, and if there is a single value of  $m$  suitable for such conditions, it is likely to be between 3 and 5.

Such a departure from the normal law is serious. We have seen that if  $m < 2.5$  the usual rule for estimating the uncertainty of the standard error breaks down altogether, and such values are not out of the question. We have therefore two problems. First, since enormous numbers of observations have been reduced assuming the normal law (or different hypotheses that imply it), we need a means of reassessing the accuracy of the summaries. Secondly, it is unusual for a set of observations to be sufficiently numerous to give a useful determination of  $m$  by itself; but if we assume a general value of  $m$  we can frame a general rule for dealing with even short runs by maximum likelihood and accordingly making an approximate adjustment of the  $t$  rule.

If we take 
$$\lambda = \bar{x} \pm \left\{ \frac{\sum (x - \bar{x})^2}{n(n-1)} \right\}^{1/2},$$

the uncertainty of the error term can be estimated roughly by using expectations. If  $\mu_2$  and  $\mu_4$  are the second and fourth moments of the law, we have for Type VII

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = 3 \frac{m - \frac{3}{2}}{m - \frac{5}{2}},$$

which is 5 for  $m = 4$ , while it is 3 for  $m$  infinite. Also

$$\sigma^2(\mu_2) = \frac{\mu_4}{n} - \frac{n-3}{n(n-1)} \mu_2^2 = \frac{\mu_2^2}{n} \left( \beta_2 - \frac{n-3}{n-1} \right).$$

For the normal law this is  $2\mu_2^2/(n-1)$ . For  $m = 4$  it is nearly  $4\mu_2^2/(n-1)$ . Hence if the mean and the mean-square deviation are used as estimates, and  $m = 4$ , the probability of error will approximately follow a  $t$  rule with  $\frac{1}{2}(n-1)$  degrees of freedom instead of  $n-1$ .

If we take  $m = 4$  and estimate  $\lambda$  and  $\sigma$  by maximum likelihood, using

the equations 4.31 (10) and (11), it is convenient to have a table of the quantity  $w$  defined by  $w^{-1} = 1 + (x - \lambda)^2 / 2M\sigma^2$  as a function of  $(x - \lambda)/\sigma$ .

$(x - \lambda)/\sigma$	$w$	$(x - \lambda)/\sigma$	$w$	$(x - \lambda)/\sigma$	$w$
0	1.000	2.4	0.482	4.8	0.189
0.1	0.998	2.5	0.462	4.9	0.183
0.2	0.993	2.6	0.442	5.0	0.177
0.3	0.983	2.7	0.424	5.1	0.171
0.4	0.970	2.8	0.406	5.2	0.165
0.5	0.955	2.9	0.389	5.3	0.160
0.6	0.937	3.0	0.373	5.4	0.155
0.7	0.917	3.1	0.358	5.5	0.150
0.8	0.894	3.2	0.344	5.6	0.146
0.9	0.869	3.3	0.330	5.7	0.141
1.0	0.843	3.4	0.317	5.8	0.137
1.1	0.816	3.5	0.305	5.9	0.133
1.2	0.788	3.6	0.293	6.0	0.130
1.3	0.760	3.7	0.282	6.1	0.128
1.4	0.732	3.8	0.271	6.2	0.122
1.5	0.705	3.9	0.261	6.3	0.119
1.6	0.677	4.0	0.251	6.4	0.116
1.7	0.650	4.1	0.242	6.5	0.112
1.8	0.623	4.2	0.233	6.6	0.109
1.9	0.598	4.3	0.225	6.7	0.106
2.0	0.573	4.4	0.217	6.8	0.104
2.1	0.549	4.5	0.209	6.9	0.101
2.2	0.525	4.6	0.202	7.0	0.099
2.3	0.503	4.7	0.195		

Also  $M = 2.6797$ ,  $m/M = 1.49$ .

There is no harm in practice in rounding the factors  $w$  to two figures.

Chauvenet† records a set of residuals of the measured semidiameter of Venus, in connexion with the problem of rejecting observations. Arranged in order of magnitude they are, in seconds of arc:

<i>Residual</i>	$w$
-1.40 . . . .	0.5
-0.44 . . . .	0.9
-0.30 . . . .	1.0
-0.24 . . . .	1.0
-0.22 . . . .	1.0
-0.13 . . . .	1.0
-0.05 . . . .	1.0
+0.06 . . . .	1.0
+0.10 . . . .	1.0
+0.18 . . . .	1.0
+0.20 . . . .	1.0
+0.39 . . . .	0.9
+0.48 . . . .	0.9
+0.63 . . . .	0.8
+1.01 . . . .	0.6
<hr/>	
	13.6

† *Spherical and Practical Astronomy*, 2, 562.



A simple calculation, allowing for the fact that two unknowns have been estimated, gave  $\sigma = 0.572''$ . This suggests the set of values  $w$ . With these the estimate of  $\lambda$  is  $+0.03''$ , which we may ignore, and

$$\sum w(x-a)^2 = 2.73.$$

Then a second approximation to  $\sigma^2$  is

$$s^2 = \frac{1.49}{13} \times 2.73 = 0.313, \quad s = 0.559''.$$

Recomputing with this value we find that the weights are unaltered to the first decimal, and we do not need a third approximation. To find an effective number of degrees of freedom we compute the right side of 4.31 (11) with  $n = 13$ ,  $\sigma = 0.65$ ; it is 4.4, so that

$$-\frac{\partial^2}{\partial \sigma^2} \log L = \frac{4.4}{0.091} = 48; \quad n' = \frac{1}{2} 0.559^2 \times 48 = 7.5.$$

To get the uncertainty of  $a$ , put  $\lambda = +0.30$  in 4.31 (10); the sum on the right side becomes  $-2.78$ , and

$$-\frac{\partial^2}{\partial \lambda^2} \log L = \frac{1.49}{0.559^2} \frac{2.78}{0.27}.$$

Hence

$$s_\lambda = 0.143$$

and the result is  $\lambda = +0.03 \pm 0.14$ , 7 d.f., approximately.

Chauvenet's criterion led him to the rejection of the two extreme observations and to  $\sigma = 0.339$ . The resulting standard error of the mean would be 0.094. But with  $\sigma = 0.56$  there are 3 residuals out of 15 greater than  $\sigma$ , 1 greater than  $2\sigma$ . This is not unreasonable either for index 4 or for the normal law. If we reject the two extreme observations and use  $\sigma = 0.34''$ , there are 4 out of 13 greater than  $\sigma$ , none greater than  $2\sigma$ . This would not be unreasonable for the normal law. The distribution by itself provides little evidence to decide whether Chauvenet's method of rejection or the present method is more appropriate. I should say, however, from comparison with other series, that there would be a stronger case for the present method, so long as there is no reason, recorded at the time of observing, for mistrusting particular observations. Even if the extreme observations are rightly rejected, the estimate of  $\sigma$  is based on 11 degrees of freedom, and from Fisher's  $z$  table there is a 5 per cent. chance of  $\sigma$  being 1.5 or more times the estimate. This is increased if observations are rejected.

According to the rough method based on the median, which is independent of the law of error, the median would be the eighth observation,  $+0.06$ , and limits corresponding to its standard error would be  $(15/4)^{1/2} = 1.9$  observations away. Interpolated, this puts the limits at  $-0.12$  and  $+0.17$ , so that the median of the law can be put at  $+0.03 \pm 0.145$ . This standard error happens to agree closely with that found for index 4.

The table of weights on p. 291 should be of use in a number of problems where there is at present no alternative to either keeping all the observations at full weight or rejecting some entirely. The fact that an error in  $m$  produces to the first order no error in either  $a$  or  $\sigma$  ensures that even if  $m$  is not 4 the hypothesis that it is will not give any serious errors. The importance of a very large residual is much reduced, but the slow variation of the weight with the size of the residual prevents the large shifts of the mean that may depend on what observations are rejected.

**5.8. Test for independence in rare events.** Here the null hypothesis is that the chance of the number of events in an interval of observation follows the Poisson rule. Two types of departure from the conditions for this rule have been considered, and both have led to the negative binomial rule. Both are somewhat artificial. On the other hand, any variation of the Poisson parameter, or any tendency of the events to occur in groups instead of independently, will tend to spread the law and make it more like the negative binomial. Among the various possibilities it has the great advantage that it can be definitely stated and involves just one new parameter. (Two simple Poisson rules superposed would involve three in all, the two for the separate laws and one for the fraction of the chance contained in one of them; and thus two new parameters.) If the data support it against the Poisson law, the latter is at any rate shown to be inadequate, and we can proceed to consider whether the negative binomial itself is satisfactory.

The Poisson law is the limit of the negative binomial when  $n \rightarrow \infty$ . There is a sufficient statistic for the parameter  $r$ , if the law is taken in the form we chose in 2.4 (13), but not for  $n$ . In a significance test, however, we are chiefly concerned with small values of the new parameter, which we can take to be  $1/n = \nu$ .

The law is

$$P(m | r', n, H) = \left( \frac{n}{n+r'} \right)^n \frac{n(n+1)\dots(n+m-1)}{m!} \left( \frac{r'}{n+r'} \right)^m. \quad (1)$$

Suppose that in a series of trials the value  $m_k$  occurs  $n_k$  times. Then

$$L = \left(\frac{n}{n+r'}\right)^{n \sum n_k} \prod \left(\frac{n(n+1)\dots(n+m_k-1)}{m_k!}\right)^{n_k} \left(\frac{r'}{n+r'}\right)^{\sum m_k n_k} \quad (2)$$

$$\begin{aligned} \frac{1}{L} \frac{\partial L}{\partial r'} &= -\frac{n \sum n_k}{n+r'} + \sum m_k n_k \left(\frac{1}{r'} - \frac{1}{n+r'}\right) \\ &= -\frac{n \sum n_k}{n+r'} + \frac{n \sum m_k n_k}{r'(n+r')}. \end{aligned} \quad (3)$$

Hence the maximum likelihood solution for  $r'$  is

$$r' = \frac{\sum m_k n_k}{\sum n_k}. \quad (4)$$

Thus the mean number of occurrences is a sufficient statistic for  $r'$ , irrespective of  $n$ ; we have already had this result in the extreme case of the Poisson law. The uncertainty of  $r'$ , however, does depend on  $n$ .

Now form  $J$  for the comparison of the above negative binomial law with the Poisson law

$$p_m = P(m | r, H) = e^{-r} r^m / m!. \quad (5)$$

If  $n$  is large we find

$$\begin{aligned} \log \frac{p'_m}{p_m} &= -n \log \left(1 + \frac{r'}{n}\right) + r + m \log \frac{r'}{r} - m \log(n+r') + \sum_{s=0}^{m-1} \log(n+s) \\ &\doteq -n \log \left(1 + \frac{r'}{n}\right) + r + m \left(\log \frac{r'}{r} - \log \frac{n+r'}{n}\right) + \frac{m(m-1)}{2n}, \end{aligned} \quad (6)$$

$$\begin{aligned} J &\doteq \sum \left\{ m \left(\log \frac{r'}{r} - \log \frac{n+r'}{n}\right) + \frac{m(m-1)}{2n} \right\} (p'_m - p_m) \\ &= (r' - r) \left(\log \frac{r'}{r} - \log \frac{n+r'}{n}\right) + \frac{1+1/n}{2n} r'^2 - \frac{r^2}{2n} \\ &\doteq \frac{(r' - r)^2}{r} + \frac{1}{2} \frac{r^2}{n^2}. \end{aligned} \quad (7)$$

Hence for large  $n$ ,  $r'$  and  $\nu$  are orthogonal parameters. This is another advantage of the form we have chosen for the negative binomial law.

As  $\nu \geq 0$  the approximate form of  $K$  given in 5.31 should be adapted to

$$K \sim \left(\frac{\pi N}{8}\right)^{1/2} \exp\left(-\frac{1}{2} \frac{\nu^2}{s_\nu^2}\right), \quad (8)$$

where  $N$  is the number of trials and estimated values are substituted for  $\nu$  and  $s_\nu$ . This form is to be used if  $\nu > s_\nu$ ; if  $|\nu| < s_\nu$ , the outside

factor is larger, tending to  $(\pi N/2)^{1/2}$  when  $\nu = 0$ . If  $\nu$  is small  $s_\nu$  should be nearly

$$s_\nu = \frac{1}{r} \sqrt{\frac{2}{N}}. \quad (9)$$

For the data on the numbers of men killed by the kick of a horse (p. 59) we find

$$N = 280, \quad r = 0.700,$$

and solving for  $\nu$  by minimum  $\chi'^2$ , taking  $r$  as given, we get

$$\nu = +0.053 \pm 0.074,$$

$$K \doteq 10 \exp(-0.26) \doteq 8.$$

The solution is rough;  $s_\nu$  as given by (9) would be about 0.12, the difference being due to the fact that the posterior probability distribution of  $\nu$  is far from normal. But in any case there is no doubt that the data confirm the Poisson rule and more detailed examination is unnecessary.

For the radioactivity data we have similarly

$$N = 2608, \quad r = 3.87, \quad \nu = -0.0866 \pm 0.0951,$$

the calculated standard error being 0.072. Then, since the estimate of  $\nu$  is negative, we use 2 instead of 8 in (8), and

$$K > 60.$$

The Poisson law is strongly confirmed.

In studies of factory accidents made by Miss E. M. Newbold,<sup>†</sup> strong departures from the Poisson rule were found, and there was a fairly good fit with the negative binomial. Two of Newbold's series, fitted by minimum  $\chi'^2$ , would correspond in the present notation to<sup>‡</sup>

$$r = 0.835 \pm 0.058, \quad n = 0.99 \pm 0.17; \quad N = 447;$$

$$r = 3.91 \pm 0.21, \quad n = 1.54 \pm 0.20; \quad N = 376.$$

In these cases  $\nu$  is several times its standard error and its posterior probability distribution should be nearly normal. Significance is obvious without calculation. But the first series gives more individuals with large numbers of accidents than the negative binomial would predict, and it seems that this law, though much better than Poisson's, is not altogether satisfactory for this series. Actually the mean number of occurrences was 0.978, which differs substantially from  $r$  as found by minimum  $\chi'^2$ , although the mean is a sufficient statistic.

**5.9. Introduction of new functions.** Suppose that a set of observations of a quantity  $y$  are made for different values of a variable  $x$ .

<sup>†</sup> *J. R. Stat. Soc.* 90, 1927, 487-647.

<sup>‡</sup> *Ann. Eugen.* 11, 1941, 108-14.

According to the null hypothesis  $g$ , the probability of  $y$  follows the same law for all values of  $x$ . According to  $g'$  the laws for  $y$  are displaced by a location parameter depending on  $x$ , for instance, a linear function of  $x$  or a harmonic function  $\alpha \sin \kappa x$ . This displacement is supposed specified except for an adjustable coefficient  $\alpha$ . We have now a complication, since the values of  $x$  may be arbitrarily chosen, and  $J$  will differ for different  $x$  even if the coefficient is the same. We therefore need to summarize the values of  $J$  into a single one.

In problems of this type the probability distribution of  $x$  may be regarded as fixed independently of the new parameter; the values of  $x$  may arise from some law that does not contain  $y$ , or they may be chosen deliberately by the experimenter. In the latter case the previous information  $H$  must be regarded as including the information that just those values of  $x$  will occur. Now suppose that the chance of a value  $x_r$  in an interval  $\delta x_r$  is  $p_r$ , and that that of  $y_r$  given  $x_r$  is  $f(x_r, \alpha, y_r) \delta y_r$ . Then for a general observation

$$P(\delta x_r, \delta y_r) = p_r f(x_r, \alpha, y_r) \delta y_r \quad (1)$$

and for the whole series

$$\begin{aligned} J &= \sum \sum \log \frac{f(x_r, \alpha + \Delta \alpha, y_r)}{f(x_r, \alpha, y_r)} p_r \{f(x_r, \alpha + \Delta \alpha, y_r) - f(x_r, \alpha, y_r)\} \delta y_r \\ &= \sum p_r J_r, \end{aligned} \quad (2)$$

where  $J_r$  is derived from the comparison of the laws for  $y_r$  given  $x_r$ .

In particular consider normal correlation, stated in terms of the regression of  $y$  on  $x$ . Applying 3.9 (15) to 2.5 (9) for given  $x$ ,  $\sigma$ ,  $\tau$  we find

$$\begin{aligned} J_x &= \frac{1}{2} \left[ \sqrt{\frac{1-\rho^2}{1-\rho'^2}} \cdot \frac{\tau}{\tau'} - \sqrt{\frac{1-\rho'^2}{1-\rho^2}} \cdot \frac{\tau'}{\tau} \right]^2 + \\ &\quad + \frac{1}{2} \left\{ \frac{1}{\tau^2(1-\rho^2)} + \frac{1}{\tau'^2(1-\rho'^2)} \right\} (\rho' \tau' - \rho \tau)^2 \frac{x^2}{\sigma^2}, \\ J &= \int \frac{1}{\sqrt{(2\pi)\sigma}} \exp\left(-\frac{x^2}{2\sigma^2}\right) J_x dx \\ &= -2 + \frac{1}{2} \frac{1-2\rho\rho'\tau/\tau'+\tau^2/\tau'^2}{1-\rho'^2} + \frac{1}{2} \frac{1-2\rho\rho'\tau'/\tau+\tau'^2/\tau^2}{1-\rho^2}. \end{aligned} \quad (3)$$

This is the case of 3.9 (38) when  $\sigma = \sigma'$ .

If all of a discrete set of values of  $x$  have an equal chance of occurring, it follows from (2) that  $J$  is the mean of the  $J_r$ . The extension to the case where the chance of  $x$  is uniformly distributed over an interval is immediate.

Now if there are  $n$  values  $x_r$ , each equally likely to occur, and we make

$nm$  observations, we shall expect that about  $m$  observations will be made of each value. It seems appropriate, in a case where all the  $x_r$  are fixed in advance, again to take the mean of the  $J_r$ . For if we form  $J$  for the whole of the observed values of  $x$ , it will be  $\sum J_r$ . If we take  $m$  observations for each value it will be  $m \sum J_r$ . If our results are to correspond as closely as possible to the case where about  $m$  observations for each  $x_r$  are expected to arise by chance we should therefore divide the latter sum by  $mn$ .

Alternatively we may argue that if the number of observed values of  $x_r$  is large and we take them in a random order, there is an equal chance of any particular  $x_r$  occurring in a given place in the order, and these chances are nearly independent. We then apply (2) directly.

The distinction is that in the first case we average  $J$  over the values of  $x$  that might occur; in the second we average it over the values of  $x$  that have actually occurred. The point, stated in other ways, has arisen in several previous discussions, and it appears that each choice is right in its proper place. In studying the variation of rainfall with latitude and longitude, for instance, we might proceed in three ways. (a) We might choose the latitudes and longitudes of the places for observation by means of a set of random numbers, and instal special rain-gauges at the places indicated. Since any place in the area could be chosen in this way, it is correct to take the average of  $J$  over the region. (b) We might deliberately set out the rain-gauges at equal intervals of latitude and longitude so as to cover the region. In this case we should take the mean of the values of  $J$  for the stations, but if the interval is small compared with the length and breadth of the region it will differ little from the mean over the whole region. (c) We might simply use the existing rain-gauges. Again we should take the mean of  $J$  for the stations. Its actual value, for given  $\alpha$ , will differ from that in (b). The stations might, for instance, all be in the southern half of the region. But we should consider the situation existing when such a method is adopted. There is no observational information for the northern half; there is a serious suggestion that the question can be settled from the southern half alone. In (a) and (b) the suggestion is that the effect is likely to be large enough to be detected from data over the whole region, but not likely to be detected from data for half of it. In fact the choice of design depends on the previous information and the difference in the value chosen for  $J$ , as a function of  $\alpha$ , expresses the same previous information. In testing the significance of a measured parallax of a star, for instance, we can and must take into account the fact that we

are observing from the Earth, not from a hypothetical planet associated with that star or from one in a remote nebula.

In physical subjects methods analogous to (b) and (c) will usually be adopted. (a) is used in some investigations relating to population statistics. It has the advantage over (c) that it randomizes systematic disturbances other than those directly considered. For instance, actual rain-gauges tend to be placed at low levels, whereas (a) and (b) would give high stations chances of being selected in accordance with the area of high land. In some problems (b) would suffer from a similar disadvantage to (a), though hardly in the present one (cf. also 4.9).

In what follows we shall follow the rule of (b) and (c) and take the summary value of  $J$  for given  $\alpha$  to be the mean of the values for the observed values of the (one or more) independent variables.

5.91. Suppose now that on  $q$  the measure of a variable  $x_r$  for given  $t_r$  follows a rule

$$P(dx_r | q, \sigma, t_r, H) = \frac{1}{\sqrt{(2\pi)}\sigma} \exp\left(-\frac{x_r^2}{2\sigma^2}\right) dx_r, \quad (1)$$

and that on  $q'$

$$P(dx_r | q', \sigma, \alpha, t_r, H) = \frac{1}{\sqrt{(2\pi)}\sigma} \exp\left[-\frac{\{x_r - \alpha f(t_r)\}^2}{2\sigma^2}\right] dx_r. \quad (2)$$

Then

$$J_r = \alpha^2 \bar{f}^2(t_r) / \sigma^2, \quad (3)$$

$$J = \alpha^2 \bar{f}^2(t_r) / \sigma^2, \quad (4)$$

where the bar indicates a mean over the observed  $t_r$ . Now in forming the likelihood for  $n$  observations we obtain the exponent

$$-\frac{1}{2\sigma^2} \sum \{x_r - \alpha f(t_r)\}^2. \quad (5)$$

Let  $a$  be the value of  $\alpha$  that makes this stationary. Evidently

$$a = \frac{\sum f(t_r) x_r}{\sum f^2(t_r)} \quad (6)$$

and (5) becomes

$$-\frac{1}{2\sigma^2} \left[ \sum f^2(t_r) (\alpha - a)^2 + \sum \{x_r - a f(t_r)\}^2 \right]. \quad (7)$$

The forms of (4) and (7) are exactly the same as in the test for whether a single true value agrees with zero; we have only to take this true value as being  $\alpha \sqrt{\{\bar{f}^2(t_r)\}}$ . Its estimate is  $a \sqrt{\{\bar{f}^2(t_r)\}}$ , and the second sum in (7) is the sum of the squares of the residuals. Consequently the whole of the tests related to the normal law of error can be adapted immediately to tests concerning the introduction of a new function to represent a series of measures.

**5.92. Allowance for old functions.** In most actual cases we have not simply to analyse a variation of measures in terms of random error and one new function. Usually it is already known that other functions with adjustable coefficients are relevant, even an additive constant being an example. These coefficients must themselves be found from the observations. We suppose that they are already known with sufficient accuracy for the effects of further changes to be linear, and that small changes in them make changes  $\alpha_s g_s(t_r)$  ( $s = 1$  to  $m$ ). The new function  $f(t)$  must not be linearly expressible in terms of the  $g_s(t)$ ; for if it was, any change made by it could be equally well expressed by changes of the  $\alpha_s$ . We can then suppose  $f(t)$  adjusted to be orthogonal with the  $g_s(t)$  by subtracting a suitable linear combination of the  $g_s(t)$ . Then the problem with regard to  $\alpha_s$  ( $s = 1$  to  $m$ ) is one of pure estimation and a factor  $\prod d\alpha_s$  must appear in the prior probabilities. Integration with regard to this will bring in factors  $(2\pi\sigma^2)^{1/2m}$  in the posterior probabilities on both  $q$  and  $q'$ , and the integration with regard to  $\sigma$  will replace the index  $-\frac{1}{2}n+1$  in 5.2 (22) by  $-\frac{1}{2}(n-m)+1 = -\frac{1}{2}(\nu-1)$  as before. But the  $n$  in the outside factor arises from the integration with respect to the new parameter and is unaltered. Hence the asymptotic formula corresponding to 5.2 (22) is

$$K \sim \sqrt{\left(\frac{\pi n}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{-1/2\nu + 1/2}.$$

As a rule  $n$  will be large compared with  $m$  and there will be little loss of accuracy in replacing  $n$  by  $\nu$  in the outside factor too.

As an example, consider the times of the  $P$  wave in seismology up to a distance of  $20^\circ$  as seen from the centre of the earth. The observations were very unevenly distributed with regard to distance; theoretical considerations showed that the expansion of the time in powers of the distance  $\Delta$  should contain a constant and terms in  $\Delta$  and  $(\Delta-1^\circ)^3$ , but no term in  $(\Delta-1^\circ)^2$ . The question was whether the observations supported a term in  $(\Delta-1^\circ)^4$ . A function  $F_4$ , given by

$$F_4 = \frac{1}{10,000}(\Delta-1)^4 - a - b\Delta - c(\Delta-1)^3,$$

$a$ ,  $b$ , and  $c$  being so chosen that  $F_4$  should be orthogonal with a constant,  $\Delta$ , and  $(\Delta-1^\circ)^3$  at the weights, was constructed. A least-squares solution based on about 384 observations gave the coefficient of  $F_4$ , in seconds, as  $-0.926 \pm 0.690$ . Here  $n = 384$  and the index is large enough for the exponential approximation to be used; we have then

$$K = \left(\frac{\pi \times 384}{2}\right)^{1/2} \exp\left(-\frac{0.926^2}{2 \times 0.690^2}\right) = 24.6 \exp(-0.9005) = 10.0.$$



The odds are therefore about 10 to 1 that the fourth power is not needed at these distances and that we should probably lose accuracy if we introduced it. (There is a change in the character of the solution about  $20^\circ$  that makes any polynomial approximation useless in ranges including that distance; hence the restriction of the solution to observations within  $20^\circ$ .) Here we have also an illustration of the principle of 1.61. There was no reason to suppose a cubic form final, its only justification being that it corresponds closely to the consequences of having one or more thin surface layers, each nearly uniform, resting on a region where the velocity increases linearly with depth. The structure of the upper layers led to the introduction of  $\Delta - 1^\circ$  in place of  $\Delta$ , and to the constant term in the time. The success of one form with three adjustable constants was then enough to show, first, that it was not in any case permissible on the data to introduce four, and hence that any other permissible formula must be one with three constants; second, that such a form, if it was to be valid, must give times agreeing closely with those given by the cubic.

### 5.93. Two sets of observations relevant to the same parameter.

It often happens that two measurable quantities  $x, y$  are related on  $q'$  in such a way that

$$x = \alpha f(t) \pm \sigma, \quad y = k\alpha g(t) \pm \tau, \quad (1)$$

where  $f(t), g(t)$  are known functions whose mean squares over the observed values are 1, and  $k$  is a known constant. For instance, in the measurement of the parallax of a star  $x, y$  may be the apparent disturbances in right ascension and declination, the theoretical values of which are the product of the unknown parallax into two known functions of the time. The mean square values of these functions need not be equal; hence if we use the form (1) the constant  $k$  will be needed. We take  $\sigma, \tau$  as known. Then

$$J = \frac{\alpha^2}{\sigma^2} + \frac{k^2 \alpha^2}{\tau^2}, \quad (2)$$

$$P(d\alpha | q'H) = \frac{1}{\pi} \frac{A d\alpha}{1 + A^2 \alpha^2}, \quad (3)$$

where 
$$A^2 = \frac{1}{\sigma^2} + \frac{k^2}{\tau^2}. \quad (4)$$

Let  $a, b$  be the maximum likelihood estimates of  $\alpha$  from the observations of  $x$  and  $y$  separately,  $s'^2$  and  $t'^2$  the mean square residuals; then

$$P(q | \theta H) \propto \exp \left\{ -\frac{n(s'^2 + a^2)}{2\sigma^2} - \frac{n(t'^2 + k^2 b^2)}{2\tau^2} \right\}, \quad (5)$$

$$P(q' | \theta H) \propto \int \exp \left\{ -\frac{ns'^2}{2\sigma^2} - \frac{nt'^2}{2\tau^2} - \frac{n(a-\alpha)^2}{2\sigma^2} - \frac{nk^2(b-\alpha)^2}{2\tau^2} \right\} \frac{1}{\pi} \frac{A d\alpha}{1+A^2\alpha^2}. \quad (6)$$

The maximum of the exponent is at

$$\alpha = \frac{a/\sigma^2 + k^2b/\tau^2}{1/\sigma^2 + k^2/\tau^2} \quad (7)$$

and, approximately,

$$P(q' | \theta H) \propto \sqrt{\left(\frac{2}{\pi n}\right)} \exp \left\{ -\frac{ns'^2}{2\sigma^2} - \frac{nt'^2}{2\tau^2} - \frac{nk^2(a-b)^2}{2(\tau^2 + k^2\sigma^2)} \right\}, \quad (8)$$

$$\begin{aligned} K &\sim \sqrt{\left(\frac{\pi n}{2}\right)} \exp \left\{ -\frac{na^2}{2\sigma^2} - \frac{nk^2b^2}{2\tau^2} + \frac{nk^2(a-b)^2}{2(\tau^2 + k^2\sigma^2)} \right\} \\ &= \sqrt{\left(\frac{\pi n}{2}\right)} \exp \left\{ -\frac{n(a\tau^2 + k^2b\sigma^2)^2}{2\sigma^2\tau^2(\tau^2 + k^2\sigma^2)} \right\} \\ &= \sqrt{\left(\frac{\pi n}{2}\right)} \exp \left\{ -\frac{(a/s_a^2 + b/s_b^2)^2}{2(1/s_a^2 + 1/s_b^2)} \right\}, \end{aligned} \quad (9)$$

where  $s_a$  and  $s_b$  are the separate standard errors.

When  $k$  is large or small the exponent reduces to  $-\frac{1}{2}nk^2b^2/\tau^2$  or  $-\frac{1}{2}na^2/\sigma^2$ , as we should expect. For intermediate values of  $k$ ,  $K$  may differ considerably according as the two estimates  $a$ ,  $b$  have the same sign or opposite signs, again as we should expect.

**5.94. Continuous departure from a uniform distribution of chance.** The chance of an event may be distributed continuously, often uniformly, over the range of a measured argument. The question may then be whether this chance departs from the distribution suggested. Thus it may be asked whether the variation of the numbers of earthquakes from year to year shows any evidence for an increase, or whether from day to day after a large main shock it departs from some simple law of chance. We consider here problems where the trial hypothesis  $q$  is that the distribution is uniform. We can then choose a linear function  $t$  of the argument  $x$ , so that  $t$  will be 0 at the lower and 1 at the upper limit. The chance of an event in a range  $dx$  is then  $dt$ , and that of  $n$  observations in specified ranges is  $\prod (dt)$ , provided that they are independent.

The alternative  $q'$  needs some care in statement. It is natural to suppose that the chance of an event in an interval  $dt$  is

$$\{1 + \alpha f(t)\} dt, \quad (1)$$

where  $f(t)$  is a given function and  $\int_0^1 f(t) dt = 0$ . This is satisfactory when  $\alpha$  is small, but if  $\alpha$  is large it no longer seems reasonable to take the

disturbance for each  $t$  as proportional to the same constant. Consider a circular disk, on which marbles are dropped, while the tray is agitated in its own plane. If the tray is horizontal the chance of a marble coming off is uniformly distributed with regard to the azimuth  $\theta$ . If it is slightly tilted in the direction  $\theta = 0$ , the chance will approximate to the above form with  $f(t) = \cos \theta$ . But with a larger tilt nearly the whole of the marbles will come off on the lower side, so that the chance on the upper side approximates to 0 and its distribution deviates completely from (1), with  $f(t) = \cos \theta$ , for any value of  $\alpha$ ; if we took  $\alpha > 1$  we should get negative chances, and with any  $\alpha \leq 1$  the chance of values of  $\theta$  between  $\frac{1}{2}\pi$  and  $\frac{3}{2}\pi$  would not be small. With still greater slopes nearly all the marbles would come off near the lowest point. Thus with an external force accurately proportional to  $\cos \theta$ , for any given slope, the resulting chance distribution may vary from a uniform one to one closely concentrated about a single value of  $\theta$ , in a way that cannot be represented even roughly by any function of the form (1).

If, however, we take in this case

$$P(d\theta | q'\alpha H) = A \exp(\alpha \cos \theta) d\theta, \quad (2)$$

where 
$$A \int_{-\pi}^{\pi} \exp(\alpha \cos \theta) d\theta = 1, \quad (3)$$

the conditions of the problem are satisfied. Negative chances are excluded, and with sufficiently large  $\alpha$  the chance can be arbitrarily closely concentrated about  $\theta = 0$ . Hence instead of (1) it seems reasonable to take

$$P(dt | q'\alpha H) = \exp\{\alpha f(t)\} \Big/ \int_0^1 \exp\{\alpha f(t)\} dt, \quad (4)$$

where  $\alpha$  may have any finite value.

Comparing with the null hypothesis  $\alpha = 0$  we see that  $J^{1/2}$  can range from  $-\infty$  to  $\infty$ , and for small  $\alpha$

$$\int_0^1 \exp\{\alpha f(t)\} dt = O(\alpha^2), \quad (5)$$

$$\begin{aligned} J &\doteq \int_0^1 \alpha f(t) \{\exp \alpha f(t) - 1\} dt \\ &\doteq \alpha^2 \overline{f^2}(t). \end{aligned} \quad (6)$$

Without loss of generality we can take

$$\overline{f^2}(t) = 1. \quad (7)$$

Then

$$\int_0^1 \exp\{\alpha f(t)\} dt \doteq (1 + \frac{1}{2}\alpha^2) \doteq \exp \frac{1}{2}\alpha^2, \quad (8)$$

$$P(q | H) = \frac{1}{2}, \quad (9)$$

$$P(q' d\alpha | H) \doteq \frac{1}{2\pi} \frac{d\alpha}{1 + \alpha^2}, \quad (10)$$

for small  $\alpha$ .

Let  $n$  observations occur in the intervals  $dt_r$ . Then over the range where the integrand is appreciable

$$P(\theta | qH) = \prod (dt_r), \quad (11)$$

$$P(\theta | q' \alpha H) \doteq \exp\{\alpha \sum f(t_r) - \frac{1}{2}n\alpha^2\} \prod (dt_r), \quad (12)$$

$$\begin{aligned} \frac{1}{K} &\doteq \frac{1}{\pi} \int_{-\infty}^{\infty} \exp\{\alpha \sum f(t_r) - \frac{1}{2}n\alpha^2\} \frac{d\alpha}{1 + \alpha^2} \\ &\doteq \left(\frac{2}{n\pi}\right)^{1/2} \exp\left[\frac{\{\sum f(t_r)\}^2}{2n}\right] \frac{1}{1 + \{\sum f(t_r)/n\}^2}. \end{aligned} \quad (13)$$

This is valid if  $\frac{1}{\sqrt{n}} \sum f(t_r)$  is not large; but then  $\sum f(t_r)/n$  will be small and the last factor will approximate to 1. Hence

$$K \sim \left(\frac{\pi n}{2}\right)^{1/2} \exp\left[-\frac{\{\sum f(t_r)\}^2}{2n}\right] \quad (14)$$

provided the estimate of  $\alpha$ , namely  $\frac{1}{n} \sum f(t_r)$ , is small.

The solution in the first edition used (1) and contained a factor  $c$  representing the range of  $\alpha$  permitted by the condition that a chance cannot be negative. This complication is rendered unnecessary by the modification (4).

**5.95.** It will be noticed in all these tests that the hypotheses, before they are tested, are reduced to laws expressing the probabilities of observable events. We distinguish between the law and its suggested explanation, if there is any—it is perfectly possible for a law to be established empirically without there being any apparent explanation, and it is also possible for the same law to have two or three different explanations. When stellar parallax was first discovered the question was whether the measured position of a star relative to stars in neighbouring directions showed only random variation or contained a systematic part with an annual period, the displacement from some standard position being related in a prescribed way to the earth's position

relative to the sun. This can be stated entirely in terms of the probabilities of observations, without further reference to the explanation by means of the possible finite distance of the star. The latter is reduced, before the test can be applied, to a suggestion of one new parameter that can be tested in the usual way. It happens here that the explanation existed before the relevant observations did; they were made to test a hypothesis. But it might well have happened that study of observations themselves revealed an annual variation of position between visually neighbouring stars, and then parallax would have been established—at first under some other name—and the theoretical explanation in terms of distance would have come later. Similarly the test of whether the universe has a finite curvature is not to be settled by ‘philosophical’ arguments claiming to show that it has or has not, but by the production of some observable result that would differ in the two cases. The systematic change of this result due to assuming a finite radius  $R$  would be the function  $f(t)$  of a test. Its coefficient would presumably be proportional to some negative power of  $R$ , but if a test should reveal such a term the result is an inductive inference that will be useful anyhow; it remains possible that there is some other explanation that has not been thought of, and there is a definite advantage in distinguishing between the result of observation and the explanation.

## VI

### SIGNIFICANCE TESTS: VARIOUS COMPLICATIONS

'What's one and one and one and one and one and one and one and one and one and one ?'

'I don't know,' said Alice, 'I lost count.'

'She can't do addition,' said the Red Queen.

LEWIS CARROLL, *Through the Looking-Glass*.

**6.0. Combination of Tests.** THE problems discussed in the last chapter are all similar in a set of respects. There is a clearly stated hypothesis  $q$  under discussion, and also an alternative  $q'$  involving one additional adjustable parameter, the possible range of whose values is restricted by the values of quantities that have a meaning even if the new parameter is not introduced. We are in the position at the outset of having no evidence to indicate whether the new parameter is needed, beyond the bare fact that it has been suggested as worth investigating; but the mere fact that we are seriously considering the possibility that it is zero may be associated with a presumption that if it is not zero it is probably small. Subject to these conditions we have shown how, with enough relevant evidence, high probabilities may be attached on the evidence, in some cases to the proposition that the new parameter is needed, in others to the proposition that it is not. Now at the start of a particular investigation one or more of these conditions may not be satisfied, and we have to consider what corrections are needed if they are not.

In the first place, we may have previous information about the values permitted on  $q'$ . This may occur in two ways. In the problem of the bias of dice, we supposed that the chance of a 5 or a 6, if the dice were biased, might be anything from 0 to 1. Now it may be said that this does not represent the actual state of knowledge, since it was already known that the bias is small. In that event we should have over-estimated the permitted range and therefore  $K$ ; the evidence against  $q$  is therefore stronger than the test has shown. Now there is something in this objection; but we notice that it still implies that the test has given the right answer, perhaps not as forcibly as it might, but quite forcibly enough. The difficulty about using previous information of this kind, however, is that it belongs to the category of imperfectly catalogued information that will make any quantitative theory of actual belief impossible until the phenomena of memory themselves become the subject-matter of a quantitative science; and even if this ever

happens it is possible that the use of such data will be entirely in the study of memory and not in, for instance, saying whether dice have a bias. However, all that we could say from general observation of dice, without actually keeping a record, is that all faces have sometimes occurred; we could not state the frequency of a 5 or a 6 more closely than that it is unlikely to have been under 0.1 or over 0.5. Such information would be quite useless when the question is whether the chance is  $\frac{1}{6}$  or 0.3377; and it may as well be rejected altogether. Vague information is never of much use, and it is of no use at all in testing small effects.

The matter becomes clearer on considering the following problem. Suppose that we take a sample of  $n$  to test an even chance. The approximate formula 5.1 (9) is

$$K = (2n/\pi)^{1/2} \exp(-\frac{1}{2}\chi^2). \quad (1)$$

Now suppose that we have a sample of 1,000 and that the departure makes  $K$  less than 1. If we divide the data into 9 groups and test each separately the outside factor for each is divided by 3; but at the same time we multiply all the standard errors by 3 and divide the contribution to  $\chi^2$  from a given genuine departure by 9. Thus a departure that would be shown by a sample of 1,000 may not be shown by any one of its sections. It might be said, therefore, that each section provides evidence for an even chance; therefore the whole provides evidence for an even chance; and that we have an inconsistency. This arises from an insufficient analysis of the alternative  $q'$ . The hypothesis  $q$  is a definitely stated hypothesis, leading to definite inferences.  $q'$  is not, because it contains an unknown parameter,<sup>†</sup> which we have denoted by  $p'$ , and would be  $\frac{1}{2}$  on  $q$  but might be anything from 0 to 1 on  $q'$ . Anything that alters the prior probability of  $p'$  will alter the inferences given by  $q'$ . Now the first sub-sample does alter it. We may start with probability  $\frac{1}{2}$  concentrated at  $p = \frac{1}{2}$  and the other  $\frac{1}{2}$  spread from 0 to 1. In general the first sub-sample will alter this ratio and may increase the probability that  $p = \frac{1}{2}$ ; but it also greatly changes the distribution of the probability of  $p'$  given  $q'$ , which will now be nearly normal about the sampling ratio with an assigned standard error estimated from the first sample. It is from this state of things that we start when we make our second sub-sample, not from a uniform distribution on  $q'$ . The permitted range has been cut down, effectively, to something of the order of the standard error of the sampling ratio given by the first sample. Consequently the outside factor in (1) is greatly reduced,

<sup>†</sup> This distinction appears also in Fisher's theory: see *The Design of Experiments*, 1935, p. 19.

and the second sample may give support for  $q'$  at a much smaller value of the estimated  $p' - \frac{1}{2}$  than if it started from scratch. We cannot therefore combine tests by simply multiplying the values of  $K$ . This would assume that posterior probabilities are chances, and they are not. The prior probability when each sub-sample is considered is not the original prior probability, but the posterior probability left by the previous one. We could proceed by using the sub-samples in order in this way, but we already know by 1.5 what the answer must be. The result of successive applications of the principle of inverse probability is the same as that of applying it to the whole of the data together, using the original prior probability, which in this case is the statement of ignorance. Thus if the principle is applied correctly, the probabilities being revised at each stage in accordance with the information already available, the result will be the same as if we applied it directly to the complete sample; and the answer for this is given by (1). It follows that the way of combining significance tests is not to multiply the  $K$ 's, but to add the values of  $n$  in the outside factors and to use a  $\chi^2$  based on the values estimated for  $p'$  and its standard error from all the samples together.

In the dice problem, therefore, the information contained in, say, 1,000 previous trials, even if they had been accurately recorded, could affect the result only through (1) a change in  $n$ , which would alter  $K$  by about 1 part in 600, (2) changes in the estimated  $p'$ , about which we are not in a position to say anything except by using Weldon's sample itself as our sole data, (3) a reduction of the standard error by 1 in 600. The one useful thing that the previous experience might contain, the actual number of successes, is just the one that is not sufficiently accurately recalled to be of any use. Thus in significance tests, just as in estimation problems, we have the result that vaguely remembered previous experience can at best be treated as a mere suggestion of something worth investigating; its effect in the quantitative application is utterly negligible.

Another type of previous information restricting the possible values of a new parameter, however, is important. This is where the existence of the new parameter is suggested by some external consideration that sets limits to its magnitude. A striking illustration of this is the work of Chapman and his collaborators on the lunar tide in the atmosphere.† From dynamical considerations it appears that there should be such a tide, and that it should be associated with a variation of

† *M.N.R.A.S.* 78, 1918, 635-8; *Q.J.R. Met. Soc.* 44, 1918, 271-9.



pressure on the ground, of the order of the load due to a foot of air or 0.001 inch of mercury. Actual readings of pressure are usually made to 0.001 inch, which represents the observational error; but the actual pressure fluctuates in an irregular way over about 3 inches. Now we saw that the significance test would lead to no evidence whatever about the genuineness of an effect until the standard error had been reduced by combining numerous observations to something comparable with the permitted range, and that it could lead to no decisive result until it had been made much less than this. The problem was therefore to utilize enough observations to bring the standard error down from about an inch of mercury to considerably under 0.001 inch—requiring apparently about  $10^7$  observations. In view of the large fluctuation present and unavoidable, Chapman rounded off the last figure of the pressures recorded; but he also restricted himself to those days when the pressure at Greenwich did not vary more than 0.1 inch, so that the standard error of one observation is reduced to  $0.1/\sqrt{3}$  inch; and combined hourly values of pressure for those days over 63 years, including 6,457 suitable days. Now  $0.1/(3 \times 6457 \times 24)^{1/2} = 0.00014$ . A definite result should therefore be obtained if there are no further complications. There might well be, since consecutive hourly values of a continuous function might be highly correlated and lead to an increase of uncertainty. Special attention had also to be given to the elimination of solar effects. The final result was to reveal a lunar semidiurnal variation with an amplitude of 0.000355 inch, the significance of which is shown immediately on inspection of the mean values for different distances of the moon from the meridian.

In such a case, where the hypothesis  $q'$ , that the effect sought is not zero, itself suggests a limit to its amount, it would obviously be unfair to apply the same test as in the case of complete previous ignorance of the amount. The range in which the parameter is sought is much less and the selection to be allowed for in choosing an estimate on  $q'$  is much less drastic and therefore requires a smaller allowance.

These considerations suggest an answer to the question of how significance tests should be combined in general. It often happens that we get a series of estimates of a parameter, from different sets of data, that all have the same sign and run up in magnitude to about twice the standard error. None of them taken by itself would be significant, but when they all agree in this way one begins to wonder whether they can all be accidental; one such accident, or even two with the same sign, might pass, but six may appear too many. We have seen how to

do the combination for the test of a sampling ratio. Similar considerations will apply to measures, so long as the standard errors of one observation are the same in all series. If they differ considerably a modification is needed, since two equal departures with the same standard error may give different results in a test when one is based on a few accurate observations and the other on many rough ones. The outside factor will not be simply  $(\pi \sum n/2)^{1/2}$ , since what it really depends on is the ratio of the range of the values initially possible to the standard error of the result. The former is fixed by the smallest range indicated and therefore by the most accurate observations, and the less accurate ones have nothing to say about it. It is only when they have become numerous enough to give a standard error of the mean less than the standard error of one observation in the more accurate series that they have anything important to add. If they satisfy this condition the outside factor will be got from 5.0(10) by taking  $f(a)$  from the most accurate observations, and  $a$  and  $s$  from all the series combined.

These considerations indicate how to adapt the results of the last chapter to deal with most of the possible types of departure from the conditions considered there. One further possibility is that  $q$  and  $q'$  may not be initially equally probable. Now, in accordance with our fundamental principle that the methods must not favour one hypothesis rather than another, this can occur only if definite evidence favouring  $q$  or  $q'$  is actually produced. If there is none, they are equally probable. If there is, and it is produced, it can be combined with the new information and give a better result than either separately. This difficulty can therefore easily be dealt with, in principle. But it requires attention to a further point in relation to Bernoulli's theorem. All the assessments of prior probabilities used so far have been statements of previous ignorance. Now can they be used at all stages of knowledge? Clearly not; in the combination of samples we have already seen that to use the same prior probability at all stages, instead of taking information into account as we go on, will lead to seriously wrong results. Even in a pure estimation problem it would not be strictly correct to find the ratios of the posterior probabilities for different ranges of the parameter by using sections of the observations separately and then multiplying the results, though the difference might not be serious. If we are not to run the risk of losing essential information in our possession, we must arrange to keep account of the whole of it. This is clear enough in specific problems. But do we learn anything from study of one problem

that is relevant to the prior probabilities in a different one? It appears that we do and must; for if the prior probabilities were fixed for all problems, since there is no limit to the number of problems that may arise, the prior probabilities would lead to practical certainty about the fraction of the times when  $g$  will be true, and about the number of times that a sampling ratio will lie in a definite range. But this would almost contradict our rule 5, that we cannot say anything with certainty about experience from *a priori* considerations alone. The distinction between certainty and the kind of approximation to certainty involved in Bernoulli's theorem makes it impossible to say that this is a definite contradiction, but it appears that the statement that even such an inference as this can be made in this way is so absurd that an escape must be sought. The escape is simply that prior probabilities are not permanent; the assessments will not hold at all stages of knowledge, their function being merely to show how it can begin. It is a legitimate question, therefore, to ask what assessments should replace them in any advanced subject, allowing for previous experience in that subject. The point has been noticed by Pearson in a passage already quoted (p. 115). When melting was first studied quantitatively it would have been right to attach prior probability  $\frac{1}{2}$  (or  $\frac{1}{4}$  as suggested in 3.2 (20)) to the proposition that a given pure substance would have a fixed melting-point, or, more accurately, that variations of the observed melting-point are random variations about some fixed value. It would be ridiculous to do so now. The rule has been established for one substance, and then for many; then the possibility that it is true for all comes to be seriously considered, and giving this a prior probability  $\frac{1}{2}$  or  $\frac{1}{4}$  we get a high posterior probability that it is true for all; and it is from this situation that we now proceed.

For the elementary problem of chances, similarly, we may begin with a finite prior probability that a chance is 0 or 1; but as soon as one chance is found that is neither 0 nor 1, it leads to a revision of the estimate and to the further question, 'Are all chances equal?' which a significance test answers in the negative; and then, 'Do chances show any significant departure from a uniform distribution?' Pearson† says that 'chances lie between 0 and 1, but our experience does not indicate any tendency of actual chances to cluster round any particular value in this range. . . . Those who do not accept the hypothesis of the equal distribution of ignorance are compelled to produce definite evidence of the clustering of chances, or to drop all application of past experience

† *Phil. Mag.* 13, 1907, 366.

to the judgement of probable future statistical ratios. It is perfectly easy to form new statistical algebras with other clustering of chances.' Accepting this statement for a moment, the accurate procedure at present would be to collect determinations of chances and take the prior probabilities of 0, 1, and intermediate values in proportion to the observed frequencies. The important point in this passage is the recognition that the Bayes-Laplace assessment is not a definite statement for all time, and that previous information from similar problems is relevant to the prior probability. But the statement is incomplete because in some subjects chances do cluster. The uniform assessment might have been right in genetics at the time of Mendel's original experiment, but a modern Mendelian would be entitled to use the probabilities indicated by the observed frequencies of 0:1, 1:1, 1:3, 3:5,... ratios in interpreting his results, and in fact does so roughly. Mendel's first results rested on about 8,000 observations; some hundreds would not usually be considered enough, and this corresponds to the fact that all that is now needed is to establish a high probability for one ratio compatible with the Mendelian theory against the others that have previously occurred and a background of other ratios attributable to differences of viability. Correlations in meteorology seem to be very evenly distributed, but those between human brothers seem to collect about  $\pm 0.5$ . A chemist wanting the molecular weight of a new compound would not content himself with a statement of his own determination. He carries out a complete analysis, finds one constitution consistent with all the data, and if he wants the accurate molecular weight for any other purpose he will calculate it from the International Table of Atomic Weights. The uncertainty will be that of the calculated value, not his own. Thus previous information is habitually used and allowed for, and it is not in all subjects that the previous information is of the type considered by Pearson in the passage quoted. It is not valid to group all estimates of chances or other parameters together to derive a revision of the prior probabilities, because the grouping is known to be different in different subjects, and this is already allowed for in practice, whether explicitly or not, and perhaps more drastically than theory would indicate. Thus differences of procedure in different subjects are largely explicable in terms of differences in the nature of previous results, allowed for in a way equivalent to reassessments of the prior probabilities based on previous experience. There is no need to assume any difference in the fundamental principles, which themselves provide means of making such reassessments. It is, in fact, desirable

that the results of a subject should be analysed at convenient intervals so as to see whether any alteration will be needed for future use, in order that its inferences should represent as accurately as possible the knowledge available at the times when they are made. Any subject in its development provides the kind of information that is needed to bring its prior probabilities up to date. At present, however, we must be content with approximations, and in some subjects at any rate there seems to be no need for any immediate modification of the assessments used to express ignorance. In subjects where statistical methods have hitherto had little application they are suitable as they stand. It is clear that we cannot revise them in the same way in all subjects; experience in genetics is applicable to other problems in genetics, but not in earthquake statistics.

There is one possible objection to reassessment; if it is carried out, it will convince the expert or the person willing to believe that we have used the whole of the data and done the work correctly. It will not convince the beginner anxious to learn; he needs to see how the learning was done. We have already had some examples to the point. The data on criminality of twins on p. 238 were taken from Fisher's book, and quoted by him from Lange. Now both Lange and Fisher already knew a great deal about like and unlike twins, and it is possible that, on their data, the question of a significant difference was already answered, and the only question for them was how large it was—a pure problem of estimation. But a person that knows of the physical distinction, but has never thought before that there might be a mental one too, should be convinced on these data alone by a  $K$  of  $1/170$ . Compare with this the results of the cattle inoculation test, where  $K = 0.37$ . The odds on these data that the inoculation is useful are about the same as that we shall pick a white ball at random out of a bag containing three white and one black, or that we shall throw a head within the first two throws with a penny. The proper judgement on these data is, 'Well, there seems to be something in it, but I should want a good deal more evidence to be satisfactorily convinced.' If we say, 'Oh, but we have much more evidence', he is entitled to say, 'Why did you not produce it?' (I may say that in this case I have not the slightest idea what other evidence exists.) The best inference is always the one that takes account of the whole of the relevant evidence; but if somebody provides us with a set of data  $\theta_1$  and we take account also of additional information  $\theta_2$ , we shall obtain  $P(q | \theta_1 \theta_2 H)$ , and if we do not tell him of  $\theta_2$ , it is not his fault if he thinks we are giving him  $P(q | \theta_1 H)$  and confusion arises.

6.1. Several new parameters often arise for consideration simultaneously. This can happen in several ways. All may be independently suggested for consideration, and it merely happens that a set of observations is capable of providing answers to several independent questions, or even, in experimental work, that it has been convenient to design an experiment deliberately so as to answer them all. This is merely a slight extension of the case of one new parameter. Each parameter can be tested separately against the standard error by the usual rule. Thus in agricultural experiments the comparisons of the productivities of two varieties of crop and of the effects of two fertilizers are questions set at the start, presumably because they are worth asking, and the answer to one has nothing directly to do with the other.

In such cases we shall need a joint prior probability distribution for the two new parameters in case they may both be accepted, and consistency requires a symmetrical method. If the parameters are  $\alpha, \beta$ , we can write  $q$  for the proposition  $\alpha = \beta = 0$ ,  $q_\alpha$  for  $\alpha \neq 0, \beta = 0$ ,  $q_\beta$  for  $\alpha = 0, \beta \neq 0$ , and  $q_{\alpha\beta}$  for  $\alpha \neq 0, \beta \neq 0$ . Then it may appear that if we test  $q_\alpha$  first and then  $q_{\alpha\beta}$ , we should form  $J$  for comparison of these and use it to give a prior probability distribution for  $\beta$  given  $\alpha$ . But this leads to an inconsistency. With an obvious notation, it will not in general be true that

$$d \tan^{-1} J_\alpha^{1/2} \cdot d \tan^{-1} J_{\beta,\alpha}^{1/2} = d \tan^{-1} J_\beta^{1/2} \cdot d \tan^{-1} J_{\alpha,\beta}^{1/2}$$

so that we might be led to different results according to which of  $\alpha$  and  $\beta$  we tested first. We can obtain symmetry if we take

$$P(d\alpha d\beta | H) = \frac{1}{\pi} \frac{dJ_\alpha^{1/2}}{1+J_\alpha} \cdot \frac{1}{\pi} \frac{dJ_\beta^{1/2}}{1+J_\beta}$$

(with the usual modifications if  $J_\alpha^{1/2}$  or  $J_\beta^{1/2}$  cannot range from  $-\infty$  to  $\infty$ ). Thus  $\alpha$  and  $\beta$  are always compared with the hypothesis that both are zero.

For reasons already given (5.45) I do not think that this need for symmetry applies if  $\alpha$  is a location parameter and  $\beta$  a standard error.

6.11. A common case is where we may have to consider both whether a new function is needed and whether the standard error needs to be increased to allow for correlation between the errors. Here two parameters arise; but the test for the first may well depend on whether we accept the second. This can be treated as follows. Let  $\alpha$  be the coefficient of the new function,  $\rho$  the intraclass correlation between the observations. Then we have to compare four alternatives, since either  $\alpha$  or  $\rho$  may be 0 independently. Then let  $q$  be the proposition  $\alpha = 0, \rho = 0$ .

$q_\alpha$  is the proposition  $\alpha \neq 0, \rho = 0$ ;  $q_\rho$  is  $\alpha = 0, \rho \neq 0$ , and  $q_{\alpha\rho}$  is  $\alpha \neq 0, \rho \neq 0$ . Then we can work out as usual

$$K_\alpha = \frac{P(q|\theta H)}{P(q_\alpha|\theta H)}, \quad K_\rho = \frac{P(q|\theta H)}{P(q_\rho|\theta H)}.$$

If these are both  $> 1$ ,  $q$  is confirmed in both cases and may be retained. If one of them is  $> 1$  and the other  $< 1$ , the evidence is for the alternative that gives the latter, and against  $q$ . Thus  $q$  is disposed of and we can proceed to consider the fourth possibility. Now

$$\frac{P(q_\alpha|\theta H)}{P(q_\rho|\theta H)} = \frac{K_\rho}{K_\alpha}$$

and the more probable of the second and third alternatives is the one with the smaller  $K$ . The relevance of this parameter may then be inferred in any case. Suppose that this is  $q_\rho$ . Then we have established internal correlation and the original standard errors are irrelevant to the test of  $q_{\alpha\rho}$  against  $q_\rho$ . The comparison will therefore be in terms of the summaries by ranges or classes, not the individual observations; the standard error found for  $\alpha$  will be larger than on  $q_\alpha$ , and it is possible that  $K_\alpha$  may be less than 1 and yet that the data do not support  $\alpha$  when allowance is made for  $\rho$ . If, however,  $\alpha$  is still supported we can assert that neither  $\alpha$  nor  $\rho$  is 0. On the other hand, if  $q_\alpha$  is asserted by the first pair of tests we can still proceed to test  $\rho$ . Thus a decision between the four alternatives can always be reached.

Referring again to Weldon's dice experiment, we have an interesting illustration. The data as recorded gave the numbers of times when the 12 dice thrown at once gave 0, 1, 2, ..., 12 fives and sixes. The test for a departure of the chance from  $\frac{1}{3}$  showed that the null hypothesis must be rejected, but the evidence might conceivably arise from a non-independence of the chances for dice thrown at the same time. This was tested by Pearson by computing the expectations of the numbers of times when 0, 1, ... fives and sixes should be thrown with the revised estimate of the chance, 0.33770, and forming a new  $\chi^2$  with them. In Fisher's revision,† in which a little grouping has been done, the revised  $\chi^2$  is 8.2 on 9 degrees of freedom, so that independence may be considered satisfactorily verified and the bias accepted as the explanation of the observed departure of the sampling ratio from  $\frac{1}{3}$ .

**6.12.** Similar considerations will apply in many other cases where two or more parameters arise at once; there is a best order of procedure, which is to assert the one that is most strongly supported, reject those

† *Statistical Methods*, p. 67.

that are denied, and proceed to consider further combinations. The best way of testing differences from a systematic rule is always to arrange our work so as to ask and answer one question at a time. Thus William of Ockham's rule,<sup>†</sup> 'Entities are not to be multiplied without necessity' achieves for scientific purposes a precise and practically applicable form: *Variation is random until the contrary is shown; and new parameters in laws, when they are suggested, must be tested one at a time unless there is specific reason to the contrary.* As examples of specific reason we have the cases of two earthquake epicentres tested for identity, where, if there is a difference in latitude, there would ordinarily be one in longitude too, or of a suggested periodic variation of unknown phase, where a cosine and sine would enter for consideration together.

This rule for arranging the analysis of the data is of the first importance. We saw before that progress was possible only by testing hypotheses in turn, at each stage treating the outstanding variation as random; assuming that progress is possible we are led to the first part of the statement, and have developed means for putting it into effect, but the second has emerged from the analysis of its own accord. It is necessary to a practical development, for if it could be asked that an indefinite number of possible changes in a law should be considered simultaneously we should never be able to carry out the work at all. The charge, 'you have not considered all possible variations' is not an admissible one; the answer is, 'The onus is on you to produce *one*.' The onus of proof is always on the advocate of the more complicated hypothesis.

**6.2. Two new parameters considered simultaneously.** There are many cases where two parameters enter into a law in such a way that it would be practically meaningless to consider one without the other. The typical case is that of a periodicity. If it is present it implies the need for both a sine and a cosine. If one is needed the other will be accepted automatically as giving only a determination of phase. There may be cases where more than two parameters enter in such a way, as in the analysis of a function of position on a sphere, where all the spherical harmonics of the same degree may be taken at once.

<sup>†</sup> William of Ockham (d. 1349 ?), known as the Invincible Doctor and the Venerable Inceptor, was a remarkable man. He proved the reigning Pope guilty of seventy errors and seven heresies, and apparently died at Munich with so little attendant ceremony that there is even a doubt about the year. See the *C.D.N.B.* The above form of the principle, known as Ockham's Razor, was first given by John Ponce of Cork in 1639. Ockham and a number of contemporaries, however, had made equivalent statements. A historical treatment is given by W. M. Thorburn, *Mind*, 27, 1918, 345-53.



The simplest possible case would be the location of a point in rectangular coordinates in two dimensions, where the suggested position is the origin, and the standard errors of measures in either direction are equal. If the true coordinates on  $q'$  are  $\lambda, \mu$ , we find

$$J = (\lambda^2 + \mu^2)/\sigma^2. \quad (1)$$

Our problem is to give a prior probability distribution for  $\lambda, \mu$  given  $\sigma$ . We suppose that for given  $\lambda^2 + \mu^2$  the probability is uniformly distributed with regard to direction. Two suggestions need consideration.

We may take the probability of  $J$  to be independent of the number of new parameters; then the rule for one parameter can be taken over unchanged. Taking polar coordinates  $\rho, \phi$  we have then

$$P(d\rho | q'\sigma H) = P(dJ | q'\sigma H) = \frac{2}{\pi} \frac{dJ^{1/2}}{1+J} = \frac{2}{\pi} \frac{\sigma d\rho}{\sigma^2 + \rho^2}, \quad (2)$$

$$P(d\lambda d\mu | q'\sigma H) = \frac{2}{\pi} \frac{\sigma d\rho}{\sigma^2 + \rho^2} \frac{d\phi}{2\pi} = \frac{1}{\pi^2} \frac{\sigma d\lambda d\mu}{\rho(\sigma^2 + \rho^2)}, \quad (3)$$

since  $\rho$  can range from 0 to  $\infty$ . Integrating with regard to  $\mu$  we find

$$P(d\lambda | q'\sigma H) = \frac{1}{\pi^2} \log \frac{\sqrt{(\sigma^2 + \lambda^2)} + \sigma}{\sqrt{(\sigma^2 + \lambda^2)} - \sigma} \frac{d\lambda}{\sqrt{(\sigma^2 + \lambda^2)}}. \quad (4)$$

Alternatively we might use such a function of  $J$  that the prior probability of  $\lambda$  or  $\mu$  separately would be the same as for the introduction of one new parameter. Such a function would be

$$P(d\lambda d\mu | q'\sigma H) = \frac{\sigma}{2\pi} \frac{d\lambda d\mu}{(\sigma^2 + \rho^2)^{3/2}}. \quad (5)$$

This would lead to the consequence that the outside factor in  $K$ , for  $n$  observations, would be  $O(n)$ . This is unsatisfactory. At the worst we could test the estimate of  $\lambda$  or  $\mu$ , whichever is larger, for significance as for one new parameter and allow for selection by multiplying  $K$  by 2, and the outside factor would still be of order  $n^{1/2}$ . This would sacrifice some information, but the result should be of the right order of magnitude.

To put the matter in another way, we notice that, if  $\lambda/\sigma$  is small, (3) and (4) lead to a presumption that  $\mu$  is small too, on account of the factor  $1/\rho$ . This is entirely reasonable. If we were simply given a value of  $\lambda$  with no information about  $\mu$  except that the probability is uniformly distributed with regard to direction we should have a Cauchy law for  $\mu$ :

$$P(d\mu | \lambda H) = \frac{\lambda d\mu}{\pi(\lambda^2 + \mu^2)}.$$

But with (5), even if  $\lambda/\sigma$  is small,  $P(d\mu | q'\sigma\lambda H)$  still has a scale factor of order  $\sigma$ . That is, (3) provides a means of saying that if  $\lambda/\sigma$  is found small in an actual investigation, and we are equally prepared for any value of  $\phi$ , then  $\mu/\sigma$  is likely to be small also. (5) provides no such means.

The acceptance of (3) and therefore (4) leads to a curious consequence, namely that if measures are available of only one of  $\lambda, \mu$  the prior probability distribution for that one is appreciably different from the one we used for a single new parameter. But I think that the arguments in their favour are much stronger than this one.

We therefore adopt (3). Each observation is supposed to consist of a pair of measures  $x_r, y_r$  referring to  $\lambda, \mu$ ; we write the means as  $\bar{x}, \bar{y}$  and put

$$2ns'^2 = \sum (x_r - \bar{x})^2 + \sum (y_r - \bar{y})^2. \quad (6)$$

The analysis proceeds as follows.

$$P(q d\sigma | H) \propto d\sigma/\sigma, \quad P(q' d\sigma d\lambda d\mu | H) \propto \frac{d\sigma d\lambda d\mu}{\pi^2 \rho(\sigma^2 + \rho^2)}; \quad (7)$$

whence

$$P(q d\sigma | \theta H) \propto \frac{1}{\sigma^{2n}} \exp\left\{-\frac{2ns'^2 + n(\bar{x}^2 + \bar{y}^2)}{2\sigma^2}\right\} \frac{d\sigma}{\sigma}, \quad (8)$$

$$P(q' d\sigma d\lambda d\mu | \theta H) \propto \frac{1}{\pi^2 \sigma^{2n}} \exp\left\{-\frac{2ns'^2 + n(\lambda - \bar{x})^2 + n(\mu - \bar{y})^2}{2\sigma^2}\right\} \frac{d\sigma d\lambda d\mu}{\rho(\sigma^2 + \rho^2)}. \quad (9)$$

We are most interested in values of  $\bar{x}, \bar{y}$  appreciably greater than their standard errors, which will be about  $s'/\sqrt{n}$ , and then we can integrate (9) approximately with regard to  $\lambda$  and  $\mu$  and substitute  $\bar{x}, \bar{y}$  for them in factors raised to low powers. Then

$$P(q' d\sigma | \theta H) \propto \frac{2}{\pi n \sigma^{2n}} \exp\left(-\frac{ns'^2}{\sigma^2}\right) \frac{\sigma^2 d\sigma}{(\bar{x}^2 + \bar{y}^2)^{1/2} (\sigma^2 + \bar{x}^2 + \bar{y}^2)}, \quad (10)$$

$$K \sim \frac{n\pi}{2} \frac{(\bar{x}^2 + \bar{y}^2)^{1/2}}{s'} \left(1 + \frac{\bar{x}^2 + \bar{y}^2}{s'^2}\right) \left(1 + \frac{\bar{x}^2 + \bar{y}^2}{2s'^2}\right)^{-n}. \quad (11)$$

$n$  is already assumed fairly large. Form a generalized  $t^2$ , such that

$$\bar{x}^2 + \bar{y}^2 = t^2 s_x^2 = \frac{s'^2}{n-1} t^2, \quad (12)$$

since the number of degrees of freedom is  $2n-2$ . Then

$$K \sim \frac{n^{1/2}\pi}{2} t \left(1 + \frac{t^2}{2(n-1)}\right)^{-n+2} = \frac{n^{1/2}\pi}{2} t \left(1 + \frac{t^2}{\nu}\right)^{-1/2\nu+1}, \quad (13)$$

valid if  $t$  is more than about 2.

It is possible to reduce  $1/K$  exactly to a single integral. We have

$$\begin{aligned}\int_0^{2\pi} e^{-A \cos \phi} d\phi &= 2\pi \left\{ 1 + \sum_{m=1}^{\infty} \frac{A^{2m}}{2m!} \frac{(2m-1)(2m-3)\dots 1}{2m(2m-2)\dots 2} \right\} \\ &= 2\pi \left( 1 + \sum_{m=1}^{\infty} \frac{A^{2m}}{2^{2m} m! m!} \right).\end{aligned}\quad (14)$$

Put

$$\bar{x}^2 + \bar{y}^2 = r^2; \quad (15)$$

then

$$P(q' d\sigma d\rho | \theta H)$$

$$\begin{aligned}&\propto \frac{d\sigma d\rho}{\pi^2 \sigma^{2n}} \exp\left\{-\frac{n}{2\sigma^2}(2s'^2 + \rho^2 + r^2)\right\} \int_0^{2\pi} \exp\left(-\frac{nr\rho \cos \phi}{\sigma^2}\right) \frac{d\phi}{\sigma^2 + \rho^2} \\ &= \frac{2d\sigma d\rho}{\pi \sigma^{2n}(\sigma^2 + \rho^2)} \exp\left\{-\frac{n}{2\sigma^2}(2s'^2 + \rho^2 + r^2)\right\} \times \\ &\quad \times \left\{ 1 + \sum_{m=1}^{\infty} \left(\frac{nr\rho}{2\sigma^2}\right)^{2m} \frac{1}{m! m!} \right\}.\end{aligned}\quad (16)$$

Put now

$$\rho = \sigma v. \quad (17)$$

$$\begin{aligned}P(q' dv | \theta H) &\propto \frac{2dv}{\pi(1+v^2)} \int_0^{\infty} \frac{1}{\sigma^{2n}} \exp(-\tfrac{1}{2}nv^2) \exp\left\{-\frac{n(2s'^2 + r^2)}{2\sigma^2}\right\} \times \\ &\quad \times \left\{ 1 + \sum_{m=1}^{\infty} \left(\frac{nr v}{2\sigma}\right)^{2m} \frac{1}{m! m!} \right\} \frac{d\sigma}{\sigma}.\end{aligned}\quad (18)$$

$$\begin{aligned}\frac{1}{K} &= \frac{2}{\pi} \int_0^{\infty} \exp(-\tfrac{1}{2}nv^2) \left[ 1 + \sum_{m=1}^{\infty} \frac{n(n+1)\dots(n+m-1)}{m! m!} \left\{ \frac{n^2 r^2 v^2}{2n(2s'^2 + r^2)} \right\}^m \right] \frac{dv}{1+v^2} \\ &= \frac{2}{\pi} \int_0^{\infty} \exp(-\tfrac{1}{2}nv^2) {}_1F_1\left\{n, 1, \frac{nr^2 v^2}{2(2s'^2 + r^2)}\right\} \frac{dv}{1+v^2}\end{aligned}\quad (19)$$

$$= \frac{2}{\pi} \int_0^{\infty} \exp\left(-\frac{ns'^2 v^2}{2s'^2 + r^2}\right) {}_1F_1\left\{1-n, 1, -\frac{nr^2 v^2}{2(2s'^2 + r^2)}\right\} \frac{dv}{1+v^2}.\quad (20)$$

If  $n = 1$ ,  $r = 0$ ,  $s'$  is identically 0 and  $K$  reduces to 1 as we should expect. If  $n = 0$  it is obvious from (19) that  $K = 1$ . The resemblance to 5.2 (33) is very close.

If several old parameters have to be determined their effect is similar to that found in 5.92;  $n$  will still appear in the outside factor but will be replaced by  $\nu$  in the  $t$  factor, but in practice it will be sufficiently accurate to use  $\nu$  in both factors.

If there is a predicted standard error we shall have

$$K \sim \frac{1}{2}n^{1/2}\pi \exp(-\frac{1}{2}\chi^2) \quad (\chi > 2). \quad (21)$$

This will be applicable, in particular, when the data are frequencies.

**6.21.** Now consider the fitting of a pair of harmonics to a set of  $n$  measures of equal standard error. The law to be considered is, for the  $r$ th observation,

$$P(dx_r | \alpha, \beta, \sigma, H) = \frac{1}{\sqrt{(2\pi)\sigma}} \exp\left\{-\frac{1}{2\sigma^2}(x_r - k_r\alpha \cos t_r - k_r\beta \sin t_r)^2\right\} dx_r, \quad (22)$$

and for comparison with the law for  $\alpha = \beta = 0$

$$J_r = k_r^2(\alpha \cos t_r + \beta \sin t_r)^2/\sigma^2. \quad (23)$$

For  $n$  observations we take the mean, namely

$$J = (A\alpha^2 + 2H\alpha\beta + B\beta^2)/\sigma^2, \quad (24)$$

where

$$nA = \sum k_r^2 \cos^2 t_r, \quad nH = \sum k_r^2 \cos t_r \sin t_r, \quad nB = \sum k_r^2 \sin^2 t_r. \quad (25)$$

In practice the phase of the variation is usually initially unknown and the distribution of the observed values  $t_r$  is irrelevant to it. We need a prior probability rule for the amplitude independent of the phase. This is obtained if we take the mean of  $J$  for variations of phase  $\phi$  with  $\alpha^2 + \beta^2 = \rho^2$  kept constant; then

$$J = (1/2n) \sum k_r^2(\alpha^2 + \beta^2)/\sigma^2 = \frac{1}{2}(A+B)\rho^2/\sigma^2, \quad (26)$$

$$\begin{aligned} P(d\alpha d\beta | q'\sigma H) &= \frac{2}{\pi} \frac{dJ^{1/2}}{1+J} \frac{1}{2\pi} d\phi \\ &= \frac{(A+B)^{1/2}}{\pi^2 \sqrt{2}} \frac{\sigma d\alpha d\beta}{\sqrt{(\alpha^2 + \beta^2)\{\sigma^2 + \frac{1}{2}(A+B)(\alpha^2 + \beta^2)\}}}. \end{aligned} \quad (27)$$

We now find

$$P(q d\sigma | \theta H) \propto \sigma^{-n} \exp\left\{-\frac{n}{2\sigma^2}(s'^2 + Aa^2 + 2Hab + Bb^2)\right\} \frac{d\sigma}{\sigma}, \quad (28)$$

$P(q' d\sigma d\alpha d\beta | \theta H)$

$$\begin{aligned} &\propto \sigma^{-n} \exp\left[-\frac{n}{2\sigma^2}\{s'^2 + A(x-a)^2 + 2H(\alpha-a)(\beta-b) + B(\beta-b)^2\}\right] \times \\ &\quad \times \frac{(A+B)^{1/2}}{\pi^2 \sqrt{2}} \frac{d\sigma d\alpha d\beta}{\sqrt{(\alpha^2 + \beta^2)\{\sigma^2 + \frac{1}{2}(A+B)(\alpha^2 + \beta^2)\}}}, \end{aligned} \quad (29)$$

where  $a, b$  are the maximum likelihood estimates of  $\alpha$  and  $\beta$ . If  $\sqrt{(\alpha^2 + \beta^2)}$  is much greater than  $s'/\sqrt{n}$ , which is the important case, we can integrate approximately with regard to  $\alpha$  and  $\beta$ . Then

$$\begin{aligned} P(q' d\sigma | \theta H) &\propto \frac{\sqrt{2}}{n\pi} \left(\frac{A+B}{AB-H^2}\right)^{1/2} \sigma^{-n} \times \\ &\quad \times \exp\left(-\frac{ns'^2}{2\sigma^2}\right) \frac{\sigma^2 d\sigma}{\sqrt{(a^2 + b^2)\{\sigma^2 + \frac{1}{2}(A+B)(a^2 + b^2)\}}}. \end{aligned} \quad (30)$$

Finally, integrating with regard to  $\sigma$ , and putting  $\sigma = s$  in factors raised to low powers,

$$\frac{1}{K} \sim \frac{\sqrt{2} \left( \frac{A+B}{AB-H^2} \right)^{1/2}}{n\pi} \left( 1 + \frac{Aa^2 + 2Hab + Bb^2}{s'^2} \right)^{1/2n} \times \\ \times \frac{s}{\sqrt{(a^2+b^2)\{1+(A+B)(a^2+b^2)/2s^2\}}} \quad (31)$$

Now in finding the least squares solution we get

$$\left. \begin{aligned} \alpha + \frac{H\beta}{A} &= a + \frac{Hb}{A} \pm \frac{s}{\sqrt{nA}} \\ \beta &= b \pm \frac{s}{\sqrt{\{n(B-H^2/A)\}}} \end{aligned} \right\}, \quad (32)$$

with

$$\nu = n-2, \quad \nu s^2 = ns'^2; \quad s/\sqrt{n} = s'/\sqrt{\nu}; \quad (33)$$

$$\frac{Aa^2 + 2Hab + Bb^2}{s'^2} = n \frac{A(a + Hb/A)^2 + (B - H^2/A)b^2}{\nu s^2} \\ = \frac{1}{\nu} \left\{ \frac{(a + Hb/A)^2}{s_a^2 + Hb/A} + \frac{b^2}{s_b^2} \right\} \\ = \frac{t^2}{\nu}. \quad (34)$$

Then

$$K \sim \frac{n\pi \left( \frac{B-H^2/A}{1+B/A} \right)^{1/2}}{\sqrt{2}} \frac{\sqrt{(a^2+b^2)}}{s} \left\{ 1 + \frac{(A+B)(a^2+b^2)}{2s^2} \right\} \left( 1 + \frac{t^2}{\nu} \right)^{-1/2\nu-1}. \quad (35)$$

The calculation is simplified by the fact that  $nA$ ,  $nH$ ,  $nB$  are coefficients in the normal equations and  $n(B-H^2/A)$  is the coefficient of  $\beta$  after  $\alpha$  has been eliminated. Hence  $t^2$  is found directly from the quantities that occur in the solution and their estimated standard errors.

If  $A = B$ ,  $H = 0$ , which is a fairly common case,

$$\frac{a^2+b^2}{s^2} = \frac{1}{nA} \left( \frac{a^2}{s_a^2} + \frac{b^2}{s_b^2} \right) = \frac{2}{n(A+B)} t^2, \quad (36)$$

$$K \sim \frac{n^{1/2}\pi}{2} t \left( 1 + \frac{t^2}{\nu} \right)^{-1/2\nu}. \quad (37)$$

The approximations have supposed that  $\sqrt{(a^2+b^2)}$  is large compared with  $s/\sqrt{n}$  and small compared with  $s$ . But a further approximation has been made, as we can see by considering the case where  $H = 0$  and  $A$  is much larger than  $B$ . If  $nB$  is of order 1, and  $\alpha$  is much less than  $s$ , the variation of the exponential factor with  $\beta$  may be less rapid than

that of the factor  $(\alpha^2 + \beta^2)^{-1/2}$ . In this case all the values of  $t_r$  are near 0 or  $\pi$ . Integrating with regard to  $\beta$  in these conditions we have

$$P(q' d\sigma d\alpha | \theta H)$$

$$\propto \frac{\sigma^{-n}}{\pi^2 \sqrt{2}} \exp \left[ -\frac{n}{2\sigma^2} \{s'^2 + (\alpha - a)^2\} \right] \log \frac{\sqrt{(\sigma^2 + \frac{1}{2}\alpha^2) + \sigma}}{\sqrt{(\sigma^2 + \frac{1}{2}\alpha^2) - \sigma}} \frac{d\sigma d\alpha}{\sigma \sqrt{(\sigma^2 + \frac{1}{2}\alpha^2)}}, \quad (38)$$

$$P(q' d\sigma | \theta H) \propto \frac{\sigma^{-n}}{\pi^2 \sqrt{n}} \exp \left( -\frac{ns'^2}{2\sigma^2} \right) \log \frac{\sqrt{(\sigma^2 + \frac{1}{2}\alpha^2) + \sigma}}{\sqrt{(\sigma^2 + \frac{1}{2}\alpha^2) - \sigma}} \frac{d\sigma}{\sqrt{(\sigma^2 + \frac{1}{2}\alpha^2)}}, \quad (39)$$

$$K \sim \pi^{3/2} n^{1/2} \sqrt{\left(1 + \frac{1}{2} \frac{a^2}{s^2}\right) \left(1 + \frac{a^2}{s'^2}\right)^{-1/2n}} \log \frac{\sqrt{(s^2 + \frac{1}{2}a^2) + s}}{\sqrt{(s^2 + \frac{1}{2}a^2) - s}} \quad (40)$$

$$\doteq \frac{\pi^{3/2} n^{1/2}}{\log(8s^2/a^2)} \left(1 + \frac{t_a^2}{\nu}\right)^{-1/2\nu} \quad (41)$$

if  $a/s$  is small.

The danger signal is  $s_b > a > s_a$ . If  $n$  is large and  $a/s$  small of order  $n^{-1/2}$ , (41) may be smaller than the value given by the direct test for one unknown. The smaller value of  $K$  represents the fact that  $J$  might actually be large but that  $\alpha$  might be small owing to the observations happening to lie near  $\sin t = 0$ . We may be able to assert with some confidence that a periodic variation is present while knowing nothing about the coefficient  $\beta$  except that it is probably of the same order of magnitude as  $a$ , but might be of the order of  $s$ . The situation will of course be very unsatisfactory, but we shall have done as much as we can with the data available. The next step would be to seek for observations for such other values of  $t$  that a useful estimate of  $\beta$  can also be found, and then to apply (33).

In the case  $A = B = \frac{1}{2}$ ,  $H = 0$ , we can reduce  $1/K$  again to a single integral. The analysis is similar to that leading to 6.2 (20) and gives

$$\frac{1}{K} = \frac{2}{\pi} \int_0^\infty \exp \left( -\frac{nv^2 s'^2}{2s'^2 + a^2 + b^2} \right) {}_1F_1 \left( 1 - \frac{1}{2}n, 1, -\frac{\frac{1}{2}n(a^2 + b^2)v^2}{2s'^2 + a^2 + b^2} \right) \frac{dv}{1 + v^2}. \quad (42)$$

In the conditions stated  $n = 1$  is impossible. If  $n = 0$ ,  $K = 1$ . If  $n = 2$ ,  $s' = 0$ , and again  $K = 1$ . This is the case where there are two observations a quarter of a period apart. The result is identical with 6.2 (20) except that  $1 - \frac{1}{2}n$  replaces  $1 - n$  in the confluent hypergeometric function and  $a^2 + b^2$  replaces  $r^2$ .

There are problems where theory suggests a harmonic disturbance such as a forced oscillation with a predicted phase. We are then really testing the introduction of one new function, not two, and the rule of 5.9 applies. If the disturbance is found we can still test a displacement of phase, due for instance to damping, by a further application of 5.9 and 5.92.

Here the cosine and sine no longer enter on an equal footing because previous considerations do not make all phases equally probable on  $q'$ .

**6.22. Test of whether two laws contain a sine and cosine with the same coefficients.** This problem stands to the last in the same relation as that of 5.41 to 5.2; I shall not develop the argument in detail but proceed by analogy. A  $J$  must be defined for the difference of the two laws. It is clear that integration with regard to the differences of  $\alpha$  and  $\beta$  will bring in a factor  $n_1 n_2 / (n_1 + n_2)$  instead of  $n$ , and that the square root of this factor can be absorbed into the second factor, so that the first two factors in (35) will be replaced by

$$\frac{\pi}{\sqrt{2}} \left( \frac{n_1 n_2}{n_1 + n_2} \right)^{1/2} \left( \frac{B' - H'^2/A'}{1 + B'/A'} \right)^{1/2},$$

where  $A'$ ,  $B'$ ,  $H'$  are coefficients in the equations used to find the differences  $\alpha_2 - \alpha_1$ ,  $\beta_2 - \beta_1$ .

The determination of the corrections to a trial position of an earthquake epicentre is essentially that of determining the variation of the residuals in the times of arrival of a wave with respect to azimuth. It was found in a study of southern earthquakes† (for a different purpose) that a few pairs gave epicentres close enough to suggest identity, though they were too far apart in time for the second to be regarded as an aftershock in the usual sense. On the other hand, cases of repetition after long intervals are known, and a test of identity would be relevant to a question of whether epicentres migrate over an area. The case chosen is that of the earthquakes of 1931 February 10 and 1931 September 25. If  $x$  and  $y$  denote the angular displacements needed by the epicentre to the south and east, the trial epicentre being  $5.3^\circ$  S.,  $102.5^\circ$  E., the equations found after elimination of the time of occurrence from the normal equations were, for 1931 February 10,

$$459x + 267y = +33,$$

$$267x + 694y = -11.$$

Number of observations 30; sum of squares of residuals 108 sec.<sup>2</sup>; solution

$$x = +0.10^\circ \pm 0.10^\circ, \quad y = -0.06^\circ \pm 0.08^\circ.$$

For 1931 September 25,

$$544x + 163y = -36,$$

$$163x + 625y = +94.$$

Number of observations 35; sum of squares 202 sec.<sup>2</sup>; solution

$$x = -0.12^\circ \pm 0.10^\circ, \quad y = +0.18^\circ \pm 0.10^\circ.$$

† *M.N.R.A.S. Geophys. Suppl.* 4, 1938, 285.

The estimated standard errors of one observation are 2.0 sec. and 2.5 sec., which may be consistent and will be assumed to be. Three parameters have been estimated for each earthquake (cf. 3.52) and hence the number of degrees of freedom is  $30+35-6 = 59$ . Then

$$s^2 = (108+202)/59 = 5.25; \quad s = 2.3 \text{ sec.}$$

The question is whether the solutions indicate different values of  $x$  and  $y$  for the two earthquakes. It is best not simply to subtract the solutions because the normal equations are not orthogonal and the uncertainties of  $x$  and  $y$  are not independent. The null hypothesis is that of identity; if it was adopted we should find  $x$  and  $y$  by adding corresponding normal equations and solving. But if there is a difference we can proceed by using suffixes 1 and 2 for the two earthquakes and writing

$$x_2 = x_1 + x', \quad y_2 = y_1 + y'.$$

Then  $x'$  and  $y'$  are the new parameters whose relevance is in question. Now we notice that both sets of normal equations can be regarded as derived from a single quadratic form

$$\begin{aligned} W = & \frac{1}{2} \cdot 459x_1^2 + 267x_1y_1 + \frac{1}{2} \cdot 694y_1^2 - 33x_1 + 11y_1 + \\ & + \frac{1}{2} \cdot 544(x_1+x')^2 + 163(x_1+x')(y_1+y') + \frac{1}{2} \cdot 625(y_1+y')^2 + \\ & + 36(x_1+x') - 94(y_1+y'), \end{aligned}$$

which leads to normal equations as follows:

$$\begin{aligned} 1003x_1 + 544x' + 430y_1 + 163y' &= -3, \\ 544x_1 + 544x' + 163y_1 + 163y' &= -36, \\ 430x_1 + 163x' + 1319y_1 + 625y' &= +83, \\ 163x_1 + 163x' + 625y_1 + 625y' &= +94. \end{aligned}$$

Eliminating  $x_1$  and  $y_1$  we get

$$\begin{array}{l|l} 245x' + 108y' = -29 & \\ 108x' + 327y' = +53 & 279y' = +66 \end{array}$$

whence the solutions can be taken as

$$x' + 0.44y' = -0.12, \quad y' = +0.24,$$

the uncertainties being independent. Then

$$t^2 = \frac{245 \times 0.12^2 + 279 \times 0.24^2}{5.25} = 3.73,$$

$$x' = -0.12 - 0.44 \times 0.24 = -0.23,$$

$$\begin{aligned} K &= \frac{\pi}{\sqrt{2}} \left( \frac{35 \times 30}{65} \right)^{1/2} \left( \frac{279}{1+1.3} \right)^{1/2} \frac{\sqrt{(0.053+0.058)}}{2.3} \left( 1 + \frac{3.73}{59} \right)^{-28.5}, \text{ nearly,} \\ &= 2.2. \end{aligned}$$



The odds on the data are therefore about 2 to 1 that the epicentres were the same. The further procedure if more accuracy was required would be to drop  $x'$  and  $y'$  in the normal equations for  $x$  and  $y$ , and solve for the latter by the usual method, revising the residuals to take account of the fact that the solution will not be at the least squares solution for either separately.

The following attempt to test the annual periodicity of earthquakes is an instance of the necessity to make a thorough test of the independence of the errors before the significance of a systematic variation is established. The numbers of determinations of epicentres of earthquakes, month by month, made for the *International Seismological Summary* for the years 1918-33 were kindly supplied to me by Miss E. F. Bellamy. These do not represent the whole number of earthquakes listed; small shocks observed at only a few stations are given only in daily lists, but the list should be representative of the large and moderate shocks, for which all the observations are given in detail. As the months are unequal in length a systematic effect was first allowed for by dividing each monthly total by the ratio of the length of the month to the mean month. The resulting values were rounded to a unit, and are as follows.

	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.	Total
1918	24	40	24	27	23	34	24	36	53	30	26	31	372
1919	18	17	23	18	30	22	43	37	55	33	13	12	321
1920	33	35	17	14	32	36	24	17	58	24	21	24	335
1921	22	16	24	17	32	20	19	16	27	26	23	16	258
1922	22	23	19	32	26	31	23	32	32	17	22	31	310
1923	20	36	26	23	39	38	51	45	142	44	50	30	544
1924	34	24	46	38	45	24	42	31	84	28	34	36	466
1925	36	50	36	36	54	56	49	39	32	28	26	36	478
1926	28	27	45	29	28	55	52	114	56	75	44	56	609
1927	42	47	57	49	82	48	60	64	51	66	57	40	663
1928	36	42	62	74	61	54	41	67	41	33	38	50	599
1929	43	41	67	63	61	66	62	51	36	44	28	39	601
1930	24	37	57	44	83	41	58	40	57	80	66	66	653
1931	61	39	50	56	52	38	64	72	67	53	36	42	630
1932	36	42	42	40	50	87	43	39	47	41	40	61	568
1933	39	54	75	52	60	69	73	42	53	47	43	33	640
Total	518	570	670	612	758	719	728	742	891	669	567	603	8,047

There is on the whole a secular increase in the number per year, which is mostly due to the increase in the number of stations, many earthquakes in the first few years of the period having been presumably missed or recorded so poorly that no epicentre could be determined. We first compute  $\chi^2$  to test proportionality in the chances. It is found to be 707 on 165 degrees of freedom! No test of significance is needed. There

are four contributions of over 20: 109 for September 1923, 60 for August 1926, 25 for June 1932, and 21 for September 1924. Even apart from these extreme cases,  $\chi^2$  remains overwhelmingly large. The only years that give anything near the normal expectation are 1921, with 12.0, and 1922, with 13.4. The immediate result is that the hypothesis of independence is seriously wrong; the test has eliminated any periodicity in a year or any submultiple, and any secular change. The obvious explanation is that on an average earthquakes occur in groups of 4.3, not as separate occurrences. The enormous number in September 1923 represent aftershocks of the great Tokyo earthquake. It would be of little use to reject the years containing the very exceptional months, because the phenomenon is present, to a greater or less extent, in nearly every year.

If the residual variation from month to month was independent we might still proceed to determine a pair of Fourier coefficients, allowing for the departure from independence within a month by simply multiplying the standard error by  $4.3^{1/2} = 2.1$ . But inspection of the signs of the residuals shows that they are not independent. We can test the number of persistences and changes of sign against an even chance; but there are many small residuals and a slight oscillation among them gives numerous changes of sign and reduces the sensitiveness of the test greatly. We can recover some of the information lost in this treatment by considering only residuals over  $\pm 7$ , thus paying some attention to magnitude as well as to sign. There are 55 persistences and 34 changes, which, tested against the formula for an even chance, give  $K = 0.7$ . But the elimination of 27 parameters has introduced 27 changes of sign, and to allow for this we must reduce the number of changes by about 13. With this modification  $K$  is 0.003. Thus the lack of independence extends over more than one month, and the standard error found on this hypothesis must be multiplied by more than 2.1. The only hope is to make separate analyses for each year and examine their consistency. If  $\theta$  denotes the phase for an annual period, measured from January 16, we get the following results for the coefficients of  $\cos \theta$  and  $\sin \theta$  in the monthly numbers.

	cos	sin		cos	sin
1918	- 2.0	- 4.8	1926	- 15.8	- 18.8
1919	- 13.2	- 5.3	1927	- 8.2	+ 0.8
1920	- 2.0	- 3.5	1928	- 5.0	+ 11.3
1921	- 1.0	+ 0.2	1929	- 8.7	+ 13.8
1922	- 3.2	- 0.2	1930	- 3.7	- 5.7
1923	- 16.2	- 21.8	1931	- 7.0	- 2.5
1924	- 4.7	- 0.7	1932	- 6.0	+ 3.2
1925	- 5.5	+ 8.5	1933	- 8.7	+ 10.5

Simple means of the coefficients, with separate determinations of the standard errors, give

$$(-6.9 \pm 1.2)\cos \theta - (0.9 \pm 2.4)\sin \theta.$$

But it is very hard to see how to account for the much greater variation of the separate values for the sine than for the cosine coefficient. If we pool the two variations to get a general uncertainty the standard errors of both coefficients are 1.9, and  $t^2 = 13.3$ .  $K$  is about 0.2. This is small enough for us to say that there is substantial evidence for a periodicity, but it is not decisive. It remains possible, in fact, that a few long series of aftershocks in the summer months are responsible, in spite of the consistently negative signs of the coefficients of the cosine; though the odds are about 4 to 1 against the suggestion.

Harmonic analysis applied to the monthly totals for the whole period gives terms  $(-110.9 \pm 10.6)\cos \theta - (18.5 \pm 10.6)\sin \theta$  on the hypothesis of independence. The standard error is  $(n/72)^{1/2}$ , where  $n$  is the number of observations. Thus for one year the terms would be

$$(-6.9 \pm 0.66)\cos \theta - (1.2 \pm 0.66)\sin \theta.$$

But we know from  $\chi^2$  that the uncertainties must be multiplied by at least  $4.3^{1/2}$ , giving 1.37. The correlation between adjacent months is responsible for the rest of the increase. If it had not been for the check on independence the above determinations might have been accepted without a moment's hesitation; as it is, they may perhaps be accepted, but certainly with hesitation.

The Schuster criterion, which is frequently used to test periodicity, is really the  $\chi^2$  test adapted to two degrees of freedom. It has, however, often led to surprising results. C. G. Knott, for instance, worked out periodicities in earthquakes corresponding to various periods near a month or fortnight, some of which corresponded to some tidal effect while others did not. The amplitudes found were about twice the Schuster expectation in 7 cases out of 8.† Knott therefore expressed doubt about their genuineness. For the annual period he found (pp. 114–16) the maximum in different regions in several different months, with an excessive number in December and January, and thus just opposite to the above results.

The present analysis is not altogether satisfactory, because the list used has been subject to a certain amount of selection. Thus the Japanese (Tango) earthquake of 1927 March 7 produced 1,071 aftershocks from March 11 to June 8; of these 532 are given in the *I.S.S.*

† *Physics of Earthquake Phenomena*, 1908, 130–6.

in daily lists, but only one is treated in detail and contributes to the above totals. Most of them were small. On the other hand, some earthquakes such as the Tokyo earthquake produced long series of large aftershocks, which have contributed greatly. There are possibilities that some bias might arise in deciding which earthquakes to treat fully and which just to mention. But there seems to be no obvious way in which this could affect the instrumental records periodically, and the interval between the earthquakes and the time when the solutions were made for them has gone through all possible phases during the interval used. Yet we still have two possible explanations. Primitive earthquakes might be stimulated more readily in summer, or they might be equally likely to occur at any time of the year and tend to produce more aftershocks in summer. There is no strong theoretical reason for either hypothesis. To test them it would be necessary to have a means of identifying primitive shocks, for instance by using only earthquakes from new epicentres. Within a single series of aftershocks, that of the Tango earthquake, I have found no evidence for any failure of independence or for periodicity, the data agreeing well with a simple law of chance  $dt/(t-\alpha)$ , where  $\alpha$  is a little earlier than the time of the main shock.† If this is general the only relevant data to a periodicity would be the times of the main shocks and the number of aftershocks in each case.

Many studies of earthquake frequency do not rest on the *I.S.S.*, which is a fairly complete catalogue of the strong and moderate earthquakes, but on much less detailed lists. For instance, in a paper by S. Yamaguti,‡ which inspired me to undertake the work of 6.4, it was claimed that there was an association between the region of an earthquake and that of its predecessor, even when they were in widely different regions. His list gave only 420 earthquakes for thirty-two years; the *I.S.S.* shows that the actual number must have been about fifty times this. He was therefore not dealing with successors at all; and in three of his eight regions the excess of successors in the same region that aftershocks must have produced is replaced by a deficiency, which is presumably due to the incompleteness of the catalogue. Thus an incomplete catalogue can lead to the failure to find a genuine effect; but if any human bias enters into the selection it may easily introduce a spurious one. For these two reasons, non-randomness and possible bias in cataloguing, I have great doubts about the reality of most of the earthquake periodicities that have been claimed. (Actual examination of the

† *Gerlands Beiträge z. Geophysik*, **53**, 1938, 111–39.

‡ *Bull. Earthquake Res. Inst.*, Tokyo, **11**, 1933, 46–68.

relations between earthquakes in different regions apparently obtained by Yamaguti disclosed no apparent departure from randomness,<sup>†</sup> and the same applied to my rediscussion using the *I.S.S.*<sup>‡</sup> after the excess in the same region had been allowed for.)

**6.23. Grouping.** It has already been seen that the estimate of the uncertainty of the location parameter in an estimation problem, where the data have been grouped, is based on the standard deviation of the observations without correction for grouping. The same applies, as Fisher has again pointed out, to significance tests based on grouped data. This follows at once from the formula 5.0 (10). For the chance of getting  $\alpha$  in a given range, given  $q$  and the fact that the data have been grouped, will be given by taking (3) with the uncorrected standard error; the range of possible variation of  $\alpha$  on  $q'$  will be got by applying the grouping correction to the apparent range, thus, in the standard problem of function fitting, replacing  $s$  by  $(s^2 - \frac{1}{2}h^2)^{1/2}$  in the outside factor, which will therefore be reduced in the ratio  $(1 - h^2/12s^2)^{1/2}$ ; but this is trivial. The usual formulae should therefore be used without correction for grouping. This agrees with Fisher's recommendation.

**6.3. Partial and serial correlation.** The conditions of intraclass correlation merge into those of two still more complicated problems, those of partial correlation and serial correlation. In partial correlation an observation consists of the values of  $k$  variables,  $x_1, \dots, x_k$ , whose joint probability density on some law is proportional to  $\exp(-\frac{1}{2}W)$ , where  $W$  is a positive definite quadratic function of the  $x_s$ . The problem will be, from  $m$  such sets of observations to estimate the coefficients in  $W$ . In intraclass correlation we may regard the  $x_s$  as having independent probability distributions about a variable  $\alpha_i$ , which itself has a normal probability distribution about  $\alpha$ . Then

$$P(dx_1 \dots dx_k \mid \alpha, \sigma, \tau, H) \propto \prod dx_s \int \exp \sum \frac{(x_s - \alpha_i)^2}{2s^2} \exp \frac{(\alpha_i - \alpha)^2}{2\tau^2} d\alpha_i.$$

Integration with regard to  $\alpha_i$  gives a joint probability distribution of the form considered in partial correlation. It will, however, be symmetrical in the  $x_s$ , which is not true in general for partial correlation.

The theory of intraclass correlation assumes that the observations fall into sets, different sets being independent. There is often some reason to suppose this, but often the data occur in a definite order, and adjacent members in the order may be closely correlated. The extreme

<sup>†</sup> F. J. W. Whipple, *M.N.R.A.S. Geophys. Suppl.* **3**, 1934, 233-8.

<sup>‡</sup> *Proc. Camb. Phil. Soc.* **32**, 1936, 441-5.

case is where the observations refer to a continuous function. We might for each integral  $n$  choose  $x_n$  from a table of random numbers and then interpolate to intermediate values by one of the standard rules for numerical interpolation. The result is a continuous function and the estimated correlation between pairs of values at interval 0.1 would be nearly unity, though the original data are derived by a purely random process. Yule pointed out that many astronomical phenomena (to which may be added many meteorological ones) can be imitated by the following model. Imagine a massive pendulum of long period, slightly damped, at which a number of boys discharge pea-shooters at irregular intervals. The result will be to set the pendulum swinging in approximately its natural period  $T$ ; but the motion will be jerky. If there is a long interval when there are no hits the pendulum may come nearly to rest again and afterwards be restarted in a phase with no relation to its original one. In this problem there is a true underlying periodicity, that of a free undisturbed pendulum. But it will be quite untrue that the motion will repeat itself at regular intervals; in fact if we perform a harmonic analysis using data over too long an interval the true period may fail to reveal itself at all owing to accidental reversal of phase. What we have in fact, if we make observations at regular intervals short compared with the true period, is a strong positive correlation between consecutive values, decreasing with increasing interval, becoming negative at intervals from  $\frac{1}{4}T$  to  $\frac{3}{4}T$ , and then positive again. At sufficiently long intervals the correlation will not be significant.

In such a problem each value is highly relevant to the adjacent values, but supplementary information relative to any value can be found from others not adjacent to it, the importance of the additional information tending to zero when the interval becomes large. For a free pendulum, for instance, the displacement at one instant would be a linear function of those at the two preceding instants of observation; but if the error of observation is appreciable three adjacent observations would give a very bad determination of the period. To get the best determination from the data it will be necessary to compare observations at least a half-period apart, and it becomes a problem of great importance to decide on the best method of estimation. Much work is being done on such problems at present, though it has not yet led to a generally satisfactory theory.†

A simple rule for the invariant  $J$  can be found in a large class of cases where (1) the probability of any one observation by itself is the same

† Cf. M. G. Kendall, *Contributions to the Study of Oscillatory Time-series*, 1946.

for both laws, (2) the probability of one observation, given the law and the previous observations, depends only on the immediately preceding one. We have for the whole series,

$$J_n = \sum_r \sum \log \frac{dP(x_r | x_1 \dots x_{r-1}, \alpha', H)}{dP(x_r | x_1 \dots x_{r-1}, \alpha, H)} \times \\ \times \{P(x_1 | \alpha' H) P(x_2 | x_1 \alpha' H) \dots P(x_n | x_1 \dots x_{n-1} \alpha' H) - \\ - P(x_1 | \alpha H) P(x_2 | x_1 \alpha H) \dots P(x_n | x_1 \dots x_{n-1} \alpha H)\}.$$

The terms containing  $\log dP(x_r | \dots)$  reduce in the conditions stated to

$$\sum \log \frac{dP(x_r | x_1 \dots x_{r-1}, \alpha', H)}{dP(x_r | x_1 \dots x_{r-1}, \alpha, H)} \{P(x_1 | \alpha' H) \dots P(x_{r+1} | x_1 \dots x_r \alpha' H) - \\ - P(x_1 | \alpha H) \dots P(x_{r+1} | x_1 \dots x_r \alpha H)\} \\ = \sum \log \frac{dP(x_r | x_{r-1}, \alpha', H)}{dP(x_r | x_{r-1}, \alpha, H)} \{P(x_{r-1}, x_r, x_{r+1} | \alpha' H) - P(x_{r-1}, x_r, x_{r+1} | \alpha H)\}$$

since  $x_r$  and earlier values do not appear in the later terms in the products, which therefore add up to 1; and we can also sum over  $x_s$  for  $s < r-1$ . We can now sum over  $x_{r+1}$  and get

$$\sum \log \frac{dP(x_r | x_{r-1}, \alpha', H)}{dP(x_r | x_{r-1}, \alpha, H)} \{P(x_{r-1} | \alpha' H) P(x_r | x_{r-1}, \alpha', H) - \\ - P(x_{r-1} | \alpha H) P(x_r | x_{r-1}, \alpha, H)\}.$$

By condition (1),  $P(x_{r-1} | \alpha' H) = P(x_{r-1} | \alpha H)$ ,

and therefore this term reduces to

$$\sum P(x_{r-1} | \alpha H) J_r,$$

where

$$J_r = \sum \log \frac{dP(x_r | x_{r-1}, \alpha', H)}{dP(x_r | x_{r-1}, \alpha, H)} \{P(x_r | x_{r-1}, \alpha', H) - P(x_r | x_{r-1}, \alpha, H)\}.$$

We have to sum over the possible values of  $x_{r-1}$ , and then with regard to  $r$ . Finally, dividing by  $n$  as indicated on p. 170, we have a summary value of  $J$  which can be used as in the case of independent observations.

For  $r = 1$ ,  $J_r = 0$ ; for  $r > 1$ ,  $J_r$  is simply  $J$  for the comparison of the two laws with  $x_{r-1}$  among the data.

The simplest case of this type is where each observation is a measure and the relation between consecutive measures is of the form

$$x_r = \rho x_{r-1} \pm \tau,$$

where all  $x_r$ , taken separately, have normal probability distributions about 0 with standard error  $\sigma$ . Then

$$\tau = \sigma(1 - \rho^2)^{1/2}$$

and for different values of  $\rho$ , with  $\sigma$  fixed,  $J_r$  is the same as for com-

parison of two normal laws with true values  $\rho x_{r-1}, \rho' x_{r-1}$  and standard errors  $\sigma(1-\rho^2)^{1/2}, \sigma(1-\rho'^2)^{1/2}$ . Then  $J_r$  ( $r > 1$ ) follows from 3.9 (15):

$$J_r = \frac{1}{2} \left( \frac{\sqrt{(1-\rho'^2)}}{\sqrt{(1-\rho^2)}} - \frac{\sqrt{(1-\rho^2)}}{\sqrt{(1-\rho'^2)}} \right)^2 + \frac{1}{2\sigma^2} \left( \frac{1}{1-\rho^2} + \frac{1}{1-\rho'^2} \right) (\rho' - \rho)^2 x_{r-1}^2.$$

But 
$$P(dx_{r-1} | \sigma H) = \frac{1}{\sqrt{(2\pi)\sigma}} \exp\left(-\frac{x_{r-1}^2}{2\sigma^2}\right) dx_{r-1}.$$

Hence

$$\begin{aligned} \sum J_r &= \sum_{r=2}^n \int J_r P(dx_{r-1} | \sigma H) \\ &= \frac{1}{2}(n-1) \frac{(\rho'^2 - \rho^2)^2}{(1-\rho^2)(1-\rho'^2)} + \left( \frac{1}{1-\rho^2} + \frac{1}{1-\rho'^2} \right) (\rho' - \rho)^2 \\ &= (n-1) \frac{(\rho' - \rho)^2 (1 + \rho\rho')}{(1-\rho^2)(1-\rho'^2)}, \\ J &= \frac{1 + \rho\rho'}{(1-\rho^2)(1-\rho'^2)} (\rho' - \rho)^2. \end{aligned}$$

This is identical with  $J$  for the comparison of two correlations, the standard errors being given.

The joint likelihood for  $n$  observations is

$$\begin{aligned} &\frac{1}{(2\pi)^{1/2n} \sigma^n (1-\rho^2)^{1/2(n-1)}} \times \\ &\times \exp \left[ -\frac{x_1^2}{2\sigma^2} - \frac{1}{2\sigma^2(1-\rho^2)} \{ (x_2 - \rho x_1)^2 + \dots + (x_n - \rho x_{n-1})^2 \} \right] \prod dx_r \\ &= \frac{1}{(2\pi)^{1/2n} \sigma^n (1-\rho^2)^{1/2(n-1)}} \times \\ &\times \exp \left[ -\frac{1}{2\sigma^2(1-\rho^2)} \{ x_1^2 - 2\rho x_1 x_2 + (1+\rho^2)x_2^2 - \dots + x_n^2 \} \right] \prod dx_r. \end{aligned}$$

The interesting mathematical properties of  $J$  in this problem suggest that it might be used, but there are obvious difficulties. One is similar to what we have twice noticed already. If the suggested value of  $\rho$  is 1, and  $\rho'$  has any value other than 1,  $J$  is infinite, and the test fails. The estimation rule gives a singularity not only at  $\rho = 1$ , which might be tolerable, but also at  $-1$ , which is not. If the correlation is  $\rho$ , for values of a function taken at equal intervals, say 1, we might try to estimate  $\rho$  from observations at intervals 2. The correlation at interval 2 would be  $\rho^2$ . The same method would apply, but  $J$  would be seriously changed if we replaced  $\rho$  by  $\rho^2$  in it.



On account of the asymmetry for the first and last observations there are no sufficient statistics, but a nearly sufficient pair will be

$$s^2 = \frac{1}{n} \sum_{r=1}^n x_r^2; \quad r = \frac{\sum_{r=1}^{n-1} x_r x_{r+1}}{\frac{1}{2}(x_1^2 + x_n^2) + \sum_{r=2}^{n-1} x_r^2}.$$

This problem is given only as an illustration. In actual cases the correlation will usually run over several observations, effectively an infinite number for a continuous function, and the procedure becomes much more complicated. Further, the law itself may differ greatly from normality. I have had two cases of this myself where the problem was to estimate a predicted nearly periodic variation and the observations were affected by non-normal errors with a serial correlation between them.† A completely systematic procedure was impossible in the present state of knowledge, but approximate methods were devised that appeared fairly satisfactory in the actual problems considered.

My impression is that, though the use of  $J$  gives rules for the prior probability in many cases where they have hitherto had to be guessed, it is not of universal application. It is sufficiently successful to encourage us to hope for a general invariance rule, but not successful enough to make us think that we have yet found it. I think that the analysis of partial correlation should lead to something more satisfactory.

In problems of continuous variation with a random element the ultimate trouble is that we have not yet succeeded in stating the law properly. The most hopeful suggestion hitherto seems to be Sir G. I. Taylor's theory of diffusion by continuous movements,‡ which has been extensively used in the theory of turbulence. At least, by taking correlations between values of a variable at *any* time-interval, it avoids the need to consider a special time-interval as fundamental.

**6.4. Contingency affecting only diagonal elements.** In the simple  $2 \times 2$  contingency table we have a clear-cut test for the association of two ungraduated properties. In normal correlation we have a case where each property is measurable and the question is whether the parameter  $\rho$  is zero or not, and to provide an estimate of it if it is not. Rank correlation is an extension to the case where the properties are not necessarily measurable, but each can be arranged in a sequence of increasing intensity, and the question is whether they tend to be specially associated near one line in the diagram, usually near a diagonal

† *M.N.R.A.S.* **100**, 1940, 139–55; **102**, 1942, 194–204.

‡ *Proc. Lond. Math. Soc.* (2) **20**, 1922, 196–212.

of the table. The amounts of the displacements from this line are relevant to the question. A more extreme case is where, on the hypothesis  $q'$ , only diagonal elements would be affected. The distinction from the case of rank correlation may be illustrated by a case where the two orders are as follows:

$X$	$Y$	$X-Y$
1	2	-1
2	1	+1
3	4	-1
4	3	+1
5	6	-1
6	5	+1
7	8	-1
8	7	+1

The rank correlation is  $1-48/504 = +0.905$ . Yet not a single member occupies the same place in the two orders. We can assert a close general correspondence without there being absolute identity anywhere. But there are cases where only absolute identity is relevant to the question under test. Such a case has been discussed by W. L. Stevens,<sup>†</sup> namely that of the alleged telepathic recognition of cards. Evidence for the phenomenon would rest entirely on an excess number of cases where the presentation and identification refer to the same card; if the card presented is the king of spades, the subject is equally wrong whether he identifies it as the king of clubs, the queen of spades, or the two of diamonds. (I am not sure whether this is right, but it is part of the conditions of the problem.) Another case is the tendency of an earthquake in a region to be followed by another in the same region; to test such a tendency we cannot use rank correlation because the regions cannot be arranged in a single order. The known phenomenon is that a large earthquake is often followed by a number of others in the same neighbourhood; but to test whether this is an accidental association or not we must regard any pair not in the same region as unconnected, whether the separation is 2,000 or 20,000 km. Only successors in the same region are favourable to the suggested association, and we have to test whether the excess of successors in the same region is large enough to support the suggestion that one earthquake tends to stimulate another soon after and at a small distance.

In the earthquake problem, which may be representative of a large

<sup>†</sup> *Ann. Eugen.* 8, 1938, 238-44.

number of others, given that the last earthquake was in a particular region, the probability that the next will be in that region and stimulated by it is  $\alpha$ , which we may take to be the same for all earthquakes. On hypothesis  $q$ ,  $\alpha$  will be 0. The chance at any time that the next earthquake will be in the  $r$ th region is  $p_r$ . On the hypothesis of randomness the chance that the next will be in region  $r$  and the next but one in region  $s$  will be  $p_r p_s$ , where all the  $p$ 's will have to be found from the data. On hypothesis  $q'$ , the chance that an earthquake will be in region  $r$  and followed by one stimulated by it will be  $p_r \alpha$ , leaving  $p_r(1-\alpha)$  to be distributed in proportion to the  $p_s$  (including  $s = r$  since we are not considering on  $q'$  that the occurrence of an earthquake in a region precludes the possibility that the next will be an independent one in the same region). Thus the joint chance will be  $(1-\alpha)p_r p_s$ , except for  $s = r$ , for which it is  $(1-\alpha)p_r^2 + \alpha p_r$ . Proceeding to the third and neglecting any influence of an earthquake other than its immediate predecessor, the joint chance of all three will be obtained by multiplying these expressions by  $(1-\alpha)p_t$  if  $t \neq s$ , and by  $(1-\alpha)p_s + \alpha$  if  $t = s$ . So we may proceed. The joint chance of a set of earthquakes, in a particular order, such that in  $x_{rs}$  cases an earthquake in region  $r$  is followed by one in region  $s$ , for all values of  $r$  and  $s$ , is

$$(1-\alpha)^{N-1} \prod (p_r)^{x_r} \prod \left(1 + \frac{\alpha}{(1-\alpha)p_r}\right)^{x_{rr}}, \quad (1)$$

where 
$$x_r = \sum_s x_{rs}, \quad N = \sum_r x_r, \quad (2)$$

and the last factor is the product over all repetitions. Then this is  $P(\theta | q', p_r, \alpha, H)$ .  $P(\theta | q, p_r, H)$  is got by putting  $\alpha = 0$ .

The invariant  $J$  for comparison of  $q$  and  $q'$  can be found by the method of 6.3. We have, if the  $(m-1)$ th observation is in region  $r$ ,

$$\begin{aligned} J_m &= \sum_r \log \left\{ \frac{(1-\alpha)p_r + \alpha}{p_r} \right\} \{ (1-\alpha)p_r + \alpha - p_r \} + \\ &\quad + \sum_r \sum'_s \log(1-\alpha) \{ (1-\alpha)p_s - p_s \} \\ &= \sum_r \log \left( 1 - \alpha + \frac{\alpha}{p_r} \right) \alpha(1-p_r) - \sum_r \log(1-\alpha) \cdot \alpha(1-p_r) \\ &= \sum_r \alpha(1-p_r) \log \left( 1 + \frac{\alpha}{p_r(1-\alpha)} \right), \\ J &= \sum_r p_r \alpha(1-p_r) \log \left( 1 + \frac{\alpha}{p_r(1-\alpha)} \right) \doteq \sum_r (1-p_r) \frac{\alpha^2}{1-\alpha} = \frac{(m-1)\alpha^2}{1-\alpha}, \end{aligned}$$

where  $m$  is the number of regions.  $J$  is infinite if  $\alpha = 1$ , corresponding

to the case where, if an earthquake is in a given region, the next is certain to be in that region.  $J$  is also infinite if for some  $r$ ,

$$(1-\alpha)p_r + \alpha = 0,$$

corresponding to the case where  $\alpha$  is negative and sufficiently large numerically for the occurrence of an earthquake in some region to inhibit the occurrence of the next in that region. This might conceivably be true, since we could contemplate a state of affairs where an earthquake relieves all stress in the region and no further earthquake can occur until the stresses have had time to grow again; by which time there will almost certainly have been an earthquake somewhere else. It is therefore worth while to consider the possibility of negative  $\alpha$ . For a significance test, however, it is enough to have an approximation for  $\alpha$  small and we shall take

$$P(d\alpha | p_1 \dots p_m H) = \frac{1}{\pi} \sqrt{(m-1)} d\alpha.$$

The interpretation of the factor in  $m$  is that our way of stating the problem does not distinguish between different parts of a region. An earthquake in it may stimulate one in another part of the region, which will be reckoned as in a different region if the region is subdivided, and hence subdivision will increase the concentration of the probability of  $\alpha$  towards smaller values.

The solution is now found as usual; the factors depending on  $p_r$  are nearly the same in both  $P(q | \theta H)$  and  $P(q' | \theta H)$ , and we can substitute the approximate values

$$p_r = x_r/N$$

in the factors that also involve  $\alpha$ . Then

$$\frac{1}{K} \doteq \frac{\sqrt{(m-1)}}{\pi} \int (1-\alpha)^{N-1} \prod \left\{ 1 + \frac{N\alpha}{(1-\alpha)x_r} \right\}^{x_{rr}} d\alpha.$$

Put 
$$x_{rr} = \frac{x_r^2}{N} + x_r a_r$$

and expand the logarithm of the integrand to order  $\alpha^2$  and  $a_r \alpha$ . We find after reduction

$$\begin{aligned} \frac{1}{K} &\doteq \frac{\sqrt{(m-1)}}{\pi} \int \exp[N\alpha \sum a_r - \tfrac{1}{2}\alpha^2(m-1)N] d\alpha \\ &\doteq \frac{\sqrt{(m-1)}}{\pi} \int \exp\{-\tfrac{1}{2}(m-1)N(\alpha-a)^2 + \tfrac{1}{2}(m-1)Na^2\} d\alpha, \end{aligned}$$

where 
$$a = \frac{\sum a_r}{m-1}$$

and 
$$K \doteq \sqrt{\left(\frac{\pi N}{2}\right)} \exp\{-\tfrac{1}{2}(m-1)Na^2\}.$$

If  $K$  is small we shall have

$$\alpha = a \pm \frac{1}{\{(m-1)N\}^{1/2}}.$$

The following table was compiled from the *International Seismological Summary* from July 1926 to December 1930. The earthquakes used were divided into ten regions; eight earthquakes in Africa were ignored because they were too few to be of any use. In some cases, also, several widely different epicentres would fit the few observations available, and these also were ignored. Thus the table is limited to fairly well observed earthquakes, which are only a fraction of those that actually occur. The North Pacific in west longitude was included with North America; the Eastern North Pacific was divided between Japan (with the Loo-Choo Islands and Formosa) and the Philippines; the East Indies were included with the South Pacific; the West Indies with Central America; and the Mediterranean region and the north coast of Africa with Europe. The results are as follows:

<i>First \ Second</i>	Europe	Asia	Indian Ocean	Japan	Philippines	South Pacific	North America	Central America	South America	Atlantic	Total	$a_r$
Europe . . .	97	58	11	73	12	60	22	22	23	19	397	+0.092
Asia . . .	69	119	13	93	21	56	16	20	22	15	444	+0.098
Indian Ocean . .	10	17	8	23	4	10	5	3	6	2	88	+0.057
Japan . . .	84	90	21	179	22	82	24	36	26	26	590	+0.077
Philippines . .	8	18	4	31	33	22	5	6	8	4	139	+0.184
South Pacific . .	57	62	14	81	17	115	22	16	22	19	425	+0.107
North America . .	17	18	3	32	6	18	21	6	6	5	132	+0.108
Central America .	16	28	4	26	5	22	2	16	10	2	131	+0.072
South America . .	29	19	4	33	9	27	7	4	24	1	157	+0.092
Atlantic . . .	10	15	6	19	10	13	8	2	10	8	101	+0.041
											2604	+0.928

Here  $m = 10$ ,  $N = 2604$ ,  $\sum a_r = 0.928$ . Then

$$K = 1.6 \times 10^{-53}.$$

The evidence for  $q'$  is therefore overwhelming. The estimate of  $\alpha$  is

$$\alpha = +0.1031 \pm 0.0065.$$

This can be interpreted as the chance that a given earthquake will be followed by an aftershock, strong enough to be widely recorded, before there has been another widely recorded earthquake anywhere else.

**6.5. Deduction as an approximation.** We have seen that in significance tests enormous odds are often obtained against the null

hypothesis, but that those obtained for it are usually much smaller. A large discrepancy makes  $K$  exponentially small, but even exact agreement with the predictions made by the null hypothesis only makes  $K$  of order  $n^{1/2}$ . But a small  $K$  does not establish the hypothesis  $q'$ . It only shows that the hypothesis that one new parameter is needed, the rest of the variation being regarded as random, is more probable than that the whole variation is random. It does not say that no further parameter is still needed. Before we can actually attach a high probability to  $q'$  in its present form we must treat it as a new  $q$  and test possible departures from it; and it is only if it survives these tests that it can be used for prediction. Thus when a hypothesis comes to be actually used, on the ground that it is 'supported by the observations', the probability that it is false is always of order  $n^{-1/2}$ , which may be as large as 0.2 and will hardly ever be as small as 0.001. Strictly, therefore, any inferences that we draw from the data should not be the inferences from  $q$  alone but from  $q$  together with all the alternatives that have been considered but found not to be supported by the data, with allowance for their posterior probabilities. If, for instance,  $x$  denotes the proposition that some future observation will lie in a particular range, and we consider a set of alternative hypotheses  $q_1, q_2, \dots$ , we shall have

$$P(x | \theta H) = \sum P(q_r x | \theta H) = \sum P(x | q_r \theta H) P(q_r | \theta H).$$

Now if in a given case one of the hypotheses,  $q$  say, has a high probability on the data, and all the others correspondingly small ones,  $P(x | \theta H)$  will be high if  $x$  has a high probability on  $q$ . If  $x$  has a low probability on  $q$ , its probability will be composed of the small part from  $q$ , representing the tail of the distribution of the chance on  $q$ , and of the various contributions from the other  $q_r$ . But the last together make up  $q'$ , and the total probability of all such values cannot exceed the posterior probability of  $q'$ . Thus the total posterior probability that the observation will be in a range improbable on  $q$  will be small. In our case the situation is more extreme, for the  $q_r$  will be statements of possible values of a parameter  $\alpha$ , which we may take to be 0 on  $q$ . But when  $K$  is large nearly all the total probability of  $q'$  comes from values of  $\alpha$  near the maximum likelihood solution, which itself is small and will give therefore almost the same inferences as  $q$ . The only effect of  $q'$  is to add to the distribution on  $q$  another about nearly the same maximum and with a slightly larger scatter and a smaller total area. Thus the total distribution on data  $\theta H$  is practically the same as on  $q\theta H$  alone; the statement of  $\theta$  takes care of the uncertainties on the data of

the parameters that are relevant on  $q$ . Thus if  $q$  has been found to be supported by the data we can take as a good approximation

$$P(x|\theta H) = P(x|q\theta H),$$

thus virtually asserting  $q$  and neglecting the alternatives. We have in fact reached an instance of the theorem of 1.6, that a well-verified hypothesis will probably continue to lead to correct inferences even if it is wrong. The only alternatives not excluded by the data are those that lead to almost the same inferences as the one adopted. The difference from the inferences in a simple estimation problem is that the bulk of the probability distribution of  $\alpha$  is concentrated in  $\alpha = 0$  instead of being about the maximum likelihood solution.

This approximation means an enormous practical convenience. In theory we never dispose completely of  $q'$ , and to be exact we should allow for the contributions of all non-zero values of  $\alpha$  in all future inferences. This would be hopelessly inconvenient, and indeed there is a limit to the amount of calculation that can be undertaken at all—another imperfection of the human mind. But it turns out that we need not do so; if  $K$  has been greater than 1 for all suggested modifications of  $q$  we can proceed as if  $q$  was true. At this stage science becomes deductive. This, however, is not a virtue, and it has nothing to do with pure logic. It is merely that deduction has at last found its proper place, as a convenient approximation to induction. However, at this stage all parameters in  $q$  now acquire a permanent status (at any rate until further observation shows, if ever, that  $q$  was wrong after all). Planetary theory, for instance, involves associating with each planet a certain quantity, which remains unchanged in predicting all observations. It is convenient to give this a definite name, *mass*. This process occurs at a much more elementary stage of learning. Whenever we find a set of properties so generally associated that we can infer that they will probably be associated in future instances, we can assert their general association as an approximate rule, and it becomes worth while to form the concept of things with this set of properties and give them a name. For scientific purposes reality means just this. It is not an *a priori* notion, and does not imply philosophical reality, whatever that may mean. It is simply a practical rule of method that becomes convenient when we can replace an inductive inference approximately by a deductive one. The possibility of doing it in any particular case is based on experience. Thus deduction is to be used in a rather Pickwickian sense. It no longer claims to make inferences with certainty,

for three reasons. The law used may be wrong; even if right, it contains parameters with finite uncertainties on the data, and these contribute to the uncertainty of predictions; and the prediction itself is made with a margin of uncertainty, expressing the random error of the individual observation.

It is worth while to devote some attention to considering *how* a law, once well supported, can be wrong. A new parameter rejected by a significance test need not in fact be zero. All that we say is that on the data there is a high probability that it is. But it is perfectly possible that it is not zero but too small to have been detected with the accuracy yet attained. We have seen how such small deviations from a law may be detected by a large sample when they would appear to have been denied by any sub-sample less than a certain size, and that this is not a contradiction of our general rules. But the question is whether we can allow for it by extending the meaning of  $q$  so as to say that the new parameter is not 0 but may be anywhere in some finite range. This might guard against a certain number of inferences stated with an accuracy that further work shows not to be realized. I think, however, that it is both impossible and undesirable. It is impossible because  $q$  could not then be stated; it would need to give the actual limits of the range, and these by hypothesis are unknown. Such limits would be a sheer guess and merely introduce an arbitrariness. Further, as the number of observations increases, the accuracy of an estimate also increases, and we cannot say in advance what limit, if any, it can reach. Hence if we suggest any limit on  $q$  it is possible that with enough observations we shall get an estimate on  $q'$  that makes nearly the whole chance of  $\alpha$  lie within those limits. What should we do then?  $K$  would be in the ratio of the ranges permitted on  $q'$  and  $q$ . Should we be satisfied to take the solution as it stands, or should we set up a new  $q$  that nobody has heard of before with a smaller range? I think that the latter alternative is the one any scientist would adopt. The former would say that the estimate must be accepted whether we adopt  $q$  or  $q'$ . But it is just then that we should think that the reason we have got a doubtful value within the range on  $q$  is that we took the range too large in the first place; and the only way of guarding against such a contradiction is to take the range on  $q$  zero. If there is anything to suggest a range of possible values it should go into the statement of  $q'$ , not of  $q$ .

Possible mistakes arising from parameters already considered and rejected being in fact not zero, but small compared with the critical value, can then be corrected in due course when enough information



becomes available. If we try to guard against it in advance we are not giving the inference from the data available, but simply guessing. If  $K > 1$ , then on the data the parameter probably is zero; there is no intelligible alternative. It does not help in the least to find out that a parameter is 0.1 if we say that it may not be 0 when the estimate is  $0.5 \pm 0.5$ . All that we can say is that we cannot find out that it is not 0 until we have increased our accuracy, and this is said with sufficient emphasis by making the posterior probability of  $q$  high but not 1.

A new parameter may be conspicuous without being very highly significant, or vice versa. A 5 to 0 sample appears striking evidence at first sight, but it only gives odds of 16 to 3 against an even chance. The bias in Weldon's dice experiments is hardly noticeable on inspection, but gives odds of about 1,600 to 1. With a small number of observations we can never get a very decisive result in sampling problems, and seldom get one in measurement. But with a large number we usually get one one way or the other. This is a reason for taking many observations. But the question may arise whether anomalies that need so many observations to reveal them are worth taking into account anyhow. In Weldon's experiments the excess chance is only 0.0044, and would be less than the standard error if the number of throws in a future trial is less than about 10,000. So if we propose to throw dice fewer times than this we shall gain little by taking the bias into account. Still, many important phenomena have been revealed by just this sort of analysis of numerous observations, such as the variation of latitude and many small parallaxes in astronomy. The success of Newton was not that he explained all the variation of the observed positions of the planets, but that he explained most of it. The same applies to a great part of modern experimental physics. Where a variation is almost wholly accounted for by a new function, and the observations are reasonably numerous, it is obvious on inspection and would also pass any significance test by an enormous margin. This is why so many great advances have been made without much attention to statistical theory on the part of their makers. But when we come to deal with smaller effects an accurate analysis becomes necessary.

## VII

### FREQUENCY DEFINITIONS AND DIRECT METHODS

Lord Mansfield gave the following advice to the newly-appointed Governor of a West India Island. 'There is no difficulty in deciding a case—only hear both sides patiently, then consider what you think justice requires, and decide accordingly; but never give reasons, for your judgment will probably be right, but your reasons will certainly be wrong.'

A. H. ENGELBACH, *More Anecdotes of Bench and Bar*.

**7.0.** Most of current statistical theory, as it is stated, is made to appear to depend on one or other of various definitions of probability that claim to avoid the notion of degrees of reasonable belief. Their object is to reduce the number of postulates, a very laudable aim; if this notion could be avoided our first axiom would be unnecessary. My contention is that this axiom is necessary, and that in practice no statistician ever uses a frequency definition, but that all use the notion of degree of reasonable belief, usually without even noticing that they are using it and that by using it they are contradicting the principles they have laid down at the outset. I do not offer this as a criticism of their results. Their practice, when they come to specific applications, is mostly very good; the fault is in the precepts.

**7.01.** Three definitions have been attempted:

1. If there are  $n$  possible alternatives, for  $m$  of which  $p$  is true, then the probability of  $p$  is defined to be  $m/n$ .
2. If an event occurs a large number of times, then the probability of  $p$  is the limit of the ratio of the number of times when  $p$  will be true to the whole number of trials, when the number of trials tends to infinity.
3. An actually infinite number of possible trials is assumed. Then the probability of  $p$  is defined as the ratio of the number of cases where  $p$  is true to the whole number.

The first definition is sometimes called the 'classical' one, and is stated in much modern work, notably that of J. Neyman.† The second is the Venn limit, its chief modern exponent being R. Mises.‡ The third is the 'hypothetical infinite population', and is usually associated with the name of Fisher, though it occurred earlier in statistical mechanics in the writings of Willard Gibbs, whose 'ensemble' still plays

† *Phil. Trans. A*, **236**, 1937, 333–80.

‡ *Wahrscheinlichkeit, Statistik und Wahrheit*, 1928; *Wahrscheinlichkeitsrechnung*, 1931.

a ghostly part. The three definitions are sometimes assumed to be equivalent, but this is certainly untrue in the mathematical sense.

7.02. The first definition appears at the beginning of De Moivre's book.<sup>†</sup> It often gives a definite value to a probability; the trouble is that the value is often one that its user immediately rejects. Thus suppose that we are considering two boxes, one containing one white and one black ball, and the other one white and two black. A box is to be selected at random and then a ball at random from that box. What is the probability that the ball will be white? There are five balls, two of which are white. Therefore, according to the definition, the probability is  $\frac{2}{5}$ . But most statistical writers, including, I think, most of those that professedly accept the definition, would give  $\frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{3} = \frac{5}{12}$ . This follows at once on the present theory, the terms representing two applications of the product rule to give the probability of drawing each of the two white balls. These are then added by the addition rule. But the proposition cannot be expressed as the disjunction of 5 alternatives out of 12. My attention was called to this point by Miss J. Hosiasson.

On such a definition, again, what is the probability that the son of two dark-eyed parents will be dark-eyed? There are two possibilities, and the probability is  $\frac{1}{2}$ . A geneticist would say that if both parents had one blue-eyed parent the probability is  $\frac{3}{4}$ ; if at least one of them is homozygous it is 1. But on the definition in question, until the last possibility is definitely disproved, it remains possible that the child will be blue-eyed and there is no alternative to the assessment  $\frac{1}{2}$ . The assessment  $\frac{3}{4}$  could be obtained by the zygote theory and the definition, but then again, why should we make our definition in terms of a hypothesis about the nature of inheritance instead of the observable difference? If it is permitted to use such a hypothesis the assessment ceases to be unique, since it is now arbitrary what we are to regard as 'alternatives' for the purpose of the definition.

Similarly, the definition could attach no meaning to a statement that a die is biased. As long as no face is absolutely impossible, the probability that any particular face will appear is  $\frac{1}{6}$  and there is no more to be said.

The definition appears to give the right answer to such a question as 'What is the probability that my next hand at bridge will contain the ace of spades?' It may go to any four players and the result is  $\frac{1}{4}$ . But is the result, in this form, of the slightest use? It says nothing

<sup>†</sup> *Doctrine of Chances*, 1738.

more—in fact rather less—than that there are four possible alternatives, one of which will give me the ace of spades. If we consider the result of a particular deal as the unit ‘case’, there are  $52!/(13!)^4$  possible deals, of which  $51!/12!(13!)^3$  will give me the ace of spades. The ratio is  $\frac{1}{4}$  as before. It may appear that this gives me some help about the result of a large number of deals, but does it? There are  $\{52!/(13!)^4\}^n$  possible sets of  $n$  deals. If  $m_1$  and  $m_2$  are two integers less than  $n$ , there are

$$\sum_{m=m_1}^{m_2} \left\{ \frac{52!}{(13!)^4} \right\}^n {}^nC_m \left(\frac{1}{4}\right)^m \left(\frac{3}{4}\right)^{n-m}$$

possible sets of deals that will give me the ace from  $m_1$  to  $m_2$  times. Dividing this by the whole number of possible sets we get the binomial assessment. But on the definition the assessment means this ratio and nothing else. It does not say that I have any reason to suppose that I shall get the ace of spades between  $\frac{1}{4}n \pm \frac{1}{2}(3n)^{1/2}$  times. This can be said only if we introduce the notion of what is reasonable to expect, and say that on each occasion all deals are equally likely. If this is done the result is what we want, but unfortunately the whole object of the definition is to avoid this notion. Without it, and using only pure mathematics and ‘objectivity’, which has not been defined, I may get the ace of spades anything from 0 to  $n$  times, and there is no more to be said. Indeed, why should we not say that there are  $n+1$  possible cases, of which those from  $m_1$  to  $m_2$  are  $m_2 - m_1 + 1$ , and the probability that I shall get the ace of spades from  $m_1$  to  $m_2$  times is

$$(m_2 - m_1 + 1)/(n + 1)?$$

Either procedure would be legitimate in terms of the definition. The only reason for taking the former and not the latter is that we do consider all deals equally likely, and not all values of  $m$ . But unfortunately the users of the definition have rejected the notion of ‘equally likely’, and without it the result is ambiguous, and also useless in any case.

For continuous distributions there are an infinite number of possible cases, and the definition makes the probability, on the face of it, the ratio of two infinite numbers and therefore meaningless. Neyman and Cramér try to avoid this by considering the probability as the ratio of the measures of sets of points. But the measure of a continuous set is ambiguous until it is separately defined. If the members can be specified by associating them with the values of a continuous variable  $x$ , then they can be specified by those of any monotonic function  $f(x)$  of that variable. The theory of continuity does not specify any particular

measure, but merely that some measure exists and therefore that an infinite number of possible measures do.  $x_2 - x_1$  and  $f(x_2) - f(x_1)$  are both possible measures of the interval between two points, and are not in general in proportion. We cannot speak of the value of a probability on this definition until we have specified how the measure is to be taken. A pure mathematician, asked how to take it, would say: 'It doesn't matter; I propose to restrict myself to theorems that are true for all ways of taking it.' But unfortunately the statistician does not so restrict himself; he decides on one particular way, his theorems would be false for any other, and the reason for choosing that way is not explained. It is not even the obvious way. Where  $x$  is a continuous variable it would seem natural to take the interval between any two points as the measure, and if its range is infinite the probability for any finite range would be zero. The assessment for the normal law of error is not taken as the interval but as the integral of the law over the interval, and this integral becomes a probability, in the sense stated, only by deriving the law in a very circuitous way from the dubious hypotheses used to explain it. The measure chosen is not the only one possible, and is not the physical measure. But in modern theories of integration the measure does appear to be the physical measure; at any rate pure mathematicians are willing to consider variables with an infinite range.

Even where the definition is unambiguous, as for the cases of dice-throwing and of the offspring of two heterozygous parents, its users would not accept its results. They would proceed by stating some limit of divergence from the most probable result and rejecting the hypothesis if the divergence comes beyond this limit. In these two cases they would, in fact, accept the experimental results. But this is a contradiction. The definition is a mathematical convention involving no hypothesis at all except that a certain number of cases are possible, and the experimental results show that these cases have occurred; the hypothesis is true. Therefore the original assessment of the probability stands without alteration, and to drop it for any other value is a contradiction. Therefore I say that this definition is never used even by its advocates; it is set up and forgotten before the ink is dry. The notion that they actually use is not defined; and as the results obtained are closely in agreement with those given by the notion of reasonable degree of belief the presumption, until more evidence is available, is that this notion is used unconsciously.

Of all the theories advocated, it is the upholders of this one that

insist most on mathematical rigour, and they do, in fact, appear mostly to have a considerable command of modern mathematical technique. But when the assessments have to be made by some principle not stated in the definitions, and are often flatly contradictory to the definitions, and when the application of the final result requires an interpretation different from that given by the definitions, the claim that the elaborate use of  $\epsilon$ ,  $o(n^{-1/2})$ , and 'almost everywhere' in the intermediate stages adds anything to the rigour is on the same level as a claim that a building is strengthened by fastening a steel tie-beam into plaster at each end.

7.03. With regard to the second and third definitions, we must remember our general criteria with regard to a theory. Does it actually reduce the number of postulates, and can it be applied in practice? Now these definitions plainly do not satisfy the second criterion. No probability has ever been assessed in practice, or ever will be, by counting an infinite number of trials or finding the limit of a ratio in an infinite series. Unlike the first definition, which gave either an unacceptable assessment or numerous different assessments, these two give none at all. A definite value is got on them *only* by making a hypothesis about what the result would be. The proof even of the existence is impossible. On the limit definition, without some rule restricting the possible orders of occurrence, there might be no limit at all. The existence of the limit is taken as a postulate by Mises, whereas Venn hardly considered it as needing a postulate.† Thus there is no saving of hypotheses in any case, and the necessary existence of the limit denies the possibility of complete randomness, which would permit the ratio in an infinite series to tend to no limit. The postulate is an *a priori* statement about possible experiments and is in itself objectionable. Using the infinite population, any finite probability is the ratio of two infinite numbers and therefore is indeterminate.‡ Thus these definitions are useless for our purpose because they do not define; the existence of the quantity defined has to be taken as a postulate, and then the definitions tell us nothing about its value or its properties, which must be the subject of further postulates. From the point of

† Cf. R. Leslie Ellis, *Camb. Phil. Trans.* 8, 1849, 2. 'For myself, after giving a painful degree of attention to the point, I have been unable to sever the judgment that one event is more likely to happen than another, or that it is to be expected in preference to it, from the belief that in the long run it will occur more frequently.' Consider a biased coin, where we have no information about which way the bias is until we have experimented. At the outset neither a head nor a tail is more likely than the other at the first throw. Therefore, according to the statement, in a long series of throws heads and tails will occur equally often. This is false whichever way the bias is.

‡ W. Burnside, *Proc. Camb. Phil. Soc.* 22, 1925, 726-7; *Phil. Mag.* 1, 1926, 670-4.

view of reducing the number of postulates they give no advantage over the use of chance as a primitive notion; their only purpose is to give a meaning to chance, but they never give its actual value because the experiments contemplated in them cannot be carried out, and the existence has no practical use without the actual value. In practice those who state them do obtain quantitative results, but these are never found in terms of the definition. They are found by stating possible values or distributions of chance, applying the product and addition rules, and comparing with observations. In fact the definitions appear only at the beginning and are never heard of again, the rest of the work being done in terms of rules derivable from the notion of reasonable degree of belief; the rules cannot be proved from the definitions stated but require further postulates.

The Venn limit and the infinite population do not involve the inconsistency that is involved in the first definition when, for instance, bias of dice is asserted; since they do not specify *a priori* what the limit or the ratio must be, they make it possible to alter the estimate of it without contradiction. Venn,<sup>†</sup> considering the product rule, stated it in terms of 'cross-series'. If we consider an infinite series of propositions all entailing  $r$ ,  $P(p|r)$  and  $P(pq|r)$  would be defined by the limits of ratios in this series, but  $P(q|pr)$  requires the notion of an infinite series all implying  $p$  and  $r$ , and of a limiting ratio for the cases of  $q$  in this series. If the series used is the actual one used in assessing  $P(p|r)$ , the product rule follows by algebra; but that does not prove that all series satisfying  $p$  and  $r$  will give the same limiting ratio for  $q$ , or indeed any limit. The existence of the limit and its uniqueness must be assumed separately in every instance. Mises takes them as postulates, and the question remains whether to take them as postulates is not equivalent to denying the possibility of randomness. With the definition in terms of an infinite population the product rule cannot even be proved in the limited sense given by the Venn definition, and must be taken as a separate postulate. Thus both definitions require the existence of probabilities and the product rule to be taken as postulates, and save no hypotheses in comparison with the treatment based on the notion of degree of reasonable belief. The value of the quantity defined on them cannot be found from the definitions in any actual case. Degree of reasonable belief is at any rate accessible, and at the least it provides some justification of the product rule by pointing to a class of cases where it can be proved.

<sup>†</sup> *The Logic of Chance*, 1866, pp. 162 et seq.

It is proved in 2.13 that, in specified conditions, the limit probably exists. But this proof is in terms of the notion of degree of reasonable belief and must be rejected by anybody that rejects that notion. He must deal with the fact that in terms of the definition of randomness the ratio may tend to any limit or no limit, and must deal with it in terms of pure mathematics.

Fisher's definition becomes workable if the infinite population is replaced by a large finite population. The addition and product rules could then be proved. The difficulty that the possible ratios would depend on the number in the population would be trivial if the population is large compared with the sample; the trouble about the infinite population is that it is precisely when it becomes infinite that the ratios become indefinite. Such a definition avoids the difficulty of the De Moivre definition about the different possible ways of stating the unit alternatives. The numbers in the population would be defined as those that would be obtained, in the conditions of the experiment, in the given number of trials, and might well be unique. But there would still be some difficulties, since the actual set of observations would still have to be regarded as a random sample from the population, and the notion of 'equally probable' would enter through the notion of randomness; it is also doubtful whether this notion could be applied validly to what must in any case be the *first* sample.

**7.04.** It appears to be claimed sometimes that the three definitions are equivalent. This is not so. For dice-throwing the first gives the chance of a 5 or a 6 unambiguously as  $\frac{1}{3}$ ; but the users of all three would usually adopt the experimental result as an approximation, and it is appreciably larger—at any rate they would expect the limit in an indefinitely extended series to be more than  $\frac{1}{3}$ . The first and second definitions can be made equivalent only by assuming the existence of the limit and then treating the experimental result as irrelevant to its value. It is also sometimes stated that it is known *by experiment* that the Venn limit is identical with the ratio given by the first definition. This is simply false; and though this claim is sometimes made by good mathematicians it appears that they must have temporarily forgotten the nature of a mathematical limit. The actual number of trials is always finite, and in the mathematical sense gives no information whatever about the result of an infinite series, unless the law connecting successive terms is given; and there is no such law for random selection. It has been argued that for a finite population, sampled without replacement, the limit must be the ratio in the population. This is true, but



it gives no meaning to the statement that the ratio in  $m$  trials is likely to agree with that in the population to order  $m^{-1/2}$ . If the selection consisted of picking out all members of one type before proceeding to the other, the first statement would be true, but the second would be hopelessly wrong, and it is the second that we need for any useful theory. For sampling with replacement, even with a finite population, there is no logical proof that we shall not go on picking the same member for ever. This is relevant to the argument concerning hands at cards. The usual assessment of the chance of getting the ace  $m$  times in  $n$  deals receives an attempted justification from the fact that we should get it in just this ratio if we got each possible deal once and once only. But unfortunately the conditions refer to sampling with replacement. Long before some deals had occurred some of the earlier ones would have occurred many times, and the argument cannot be applied. The difficulty will be appreciated by those who have tried to obtain a complete set of cards, one by one, from cigarette packets each containing one. A dozen of one card may be obtained before some others have appeared at all.

Some doubt is apparently felt by the advocates of these definitions, who are liable to say when challenged on a particular mathematical point that the statement is 'reasonable'. But this gives away the entire case. *The only excuse for the definitions is that they exclude the notion of 'reasonable' in contrast to 'mathematically proved', and they therefore invite challenge on mathematical grounds. If an actual mathematical proof cannot be given, showing that a different result is simply impossible, the result is not proved. To say then that it is reasonable is mathematically meaningless, and grants that 'reasonable' has a meaning, which is indispensable to the theory, and which is neither a mathematical nor an objective meaning.* If it follows assignable rules they should be stated, which is what has been done here; if it does not, my Axiom 1 is rejected, and it is declared that it is reasonable to say, on the same data, both that  $p$  is more probable than  $q$  and  $q$  more probable than  $p$ . Curiously, however, the extreme tolerance expressed in such an attitude does not appear to be borne out in practice. The statistical journals are full of papers each maintaining, if not that the author's method is the only reasonable one, that somebody else's is not reasonable at all.

**7.05.** The most serious drawback of these definitions, however, is the deliberate omission to give any meaning to the probability of a hypothesis. All that they can do is to set up a hypothesis and give arbitrary rules for rejecting it in certain circumstances. They do not

say what hypothesis should replace it in the event of rejection, and there is no proof that the rules are the best in any sense. The scientific law is thus (apparently) made useless for purposes of inference. It is merely something set up like a coconut to stand until it is hit; an inference from it means nothing, because these treatments do not assert that there is any reason to suppose the law to be true, and it thus becomes indistinguishable from a guess. Nevertheless in practice much confidence is placed in these inferences, if not by statisticians themselves, at least by the practical men that consult them for advice. I maintain that the practical man is right; it is the statistician's agnosticism that is wrong. The statistician's attitude is, of course, opposite to that of the applied mathematician, who asserts that his laws are definitely proved. But an intermediate attitude that recognizes the validity of the notion of the probability of a law avoids both difficulties.

The actual procedure is usually independent of the definitions. A distribution of chance is set up as a hypothesis, and more complicated probabilities are derived from it by means of the addition and product rules. I have no criticism of this part of the work, since the distribution is always at the very least a suggestion worth investigation, and the two rules appear also in my theory. But the answer is necessarily in the form of a distribution of the chance of different sets of observations, given the same hypothesis. The practical problem is the inverse one; we have a unique set of observations and the problem is to decide between different hypotheses by means of it. The transition from one to the other necessarily involves some new principle. Even in pure mathematics we have this sort of ambiguity. If  $x = 1$ , it follows that  $x^2 + x - 2 = 0$ . But if  $x^2 + x - 2 = 0$ , it does not follow that  $x = 1$ . It would if we had the supplementary information that  $x$  is positive. In the probability problem the difficulty is greater, because in any use of a given set of observations to choose between different laws, or different values of parameters in the same law, we are making a selection out of a range, usually continuous, of possible values of the parameters, between which there is originally usually little to choose. (On the Venn and Fisher definitions this would mean a decision of which series or which population is to be chosen out of a super-population.) The actual selection must involve some principle that is not included in the direct treatment. The principle of inverse probability carries the transition out formally, the prior probability being chosen to express the previous information or lack of it. Rejecting the restriction of probabilities to those of observations given hypotheses and applying the rules to the

probabilities of hypotheses themselves, the principle of inverse probability is a theorem, being an immediate consequence of the product rule. No new hypothesis is needed. But the restriction spoken of makes some new hypothesis necessary, and we must examine what this is.

7.1. 'Student's' treatment of the problem of the uncertainty of the mean of a set of observations derived from the normal law provides an interesting illustration, and has the further merit of being accepted by all schools. The result actually proved is 2.8 (18)

$$P(dz | x, \sigma, H) \propto (1+z^2)^{-1/2n} dz, \quad (1)$$

where  $x$  and  $\sigma$  are the true value and standard error, supposed known, and if  $\bar{x}$  and  $s$  are the mean and standard deviation of the observations,

$$z = \frac{x - \bar{x}}{s}. \quad (2)$$

My result is, 3.41 (6),

$$P(dz | \theta H) \propto (1+z^2)^{-1/2n} dz, \quad (3)$$

which, since the right side involves the observations only through  $\bar{x}$  and  $s$ , leads, by the principle of the suppression of irrelevant data (1.7), to

$$P(dz | \bar{x}, s, H) \propto (1+z^2)^{-1/2n} dz. \quad (4)$$

This is not the same thing as (1) since the data are different. The usual way of stating (1) speaks of the probability of a proposition by itself without explicit mention of the data, and we have seen how confusing assessments on different data may lead to grossly wrong results even in very simple direct problems. In a case analogous to this we may note that the probability that Mr. Smith is dead to-day, given that he had smallpox last week, is not the same as the probability that he had smallpox last week, given that he is dead to-day. But here if we interpret (1) to mean (4) we get the correct posterior probability distribution for  $x$  given  $\bar{x}$  and  $s$ , and this is what in fact is done. But (1) certainly does not mean (4), and we must examine in what conditions it can imply it. We notice first that the inclusion of any information about  $\bar{x}$  and  $s$  in the data in (1), other than the information already given in the statement of  $x$ ,  $\sigma$ , and  $H$  (the latter involving the truth of the normal law), would make it false. For the assessment on information including the exact value of either  $\bar{x}$  or  $s$  would no longer depend on  $z$  alone, but would involve the value of  $x - \bar{x}$  or of  $s/\sigma$  explicitly. For intermediate amounts of information other parameters would appear, and would appear in the answer. Thus we cannot proceed by including  $\bar{x}$  and  $s$  in the data in (1) and then suppressing  $x$  and  $\sigma$  as irrelevant to get (4);

for if we did this the probability of  $dz$  would be unity for all ranges that included the actual value and zero for all others.

But we notice that in (1) the values of  $x$  and  $\sigma$  are irrelevant to  $z$ , and can therefore be suppressed, by Theorem 11, to give

$$P(dz | H) \propto (1+z^2)^{-1/2n} dz, \quad (5)$$

since the conditions of observation  $H$  entail the existence of  $\bar{x}$  and  $s$ ,  $x$  and  $\sigma$ , and this is the vital step. On the face of it this says nothing, for  $z$  has no value unless the quantities  $x$ ,  $\bar{x}$ , and  $s$  are given. But just for that reason it is now possible that if we now introduce  $\bar{x}$  and  $s$  into the data the form will be unaltered. The argument is apparently that the location of the probability distribution of  $x$ , given  $\bar{x}$  and  $s$ , must depend only on  $\bar{x}$ , and its scale must depend only on  $s$ . But this amounts to saying that

$$P(dz | \bar{x}, s, H) = f(z) dz; \quad (6)$$

and since  $\bar{x}$  and  $s$  are irrelevant to  $z$  they can be suppressed, and the left side reduces to  $P(dz | H)$ , which is known from (5). Thus the result (4) follows.

Something equivalent to the above seems to have been appreciated by 'Student', though it cannot be expressed in his notation. But we must notice that it involves two hypotheses: first, that nothing in the observations but  $\bar{x}$  and  $s$  is relevant; secondly, that whatever they may be in the actual observations we are at full liberty to displace or rescale the distribution in accordance with them. The first is perhaps natural, but it is desirable to keep the number of hypotheses as small as possible, whether they are natural or not, and the result is proved by the principle of inverse probability. The second can mean only one thing, that the true value  $x$  and the standard error  $\sigma$  are initially completely unknown. If we had any information about them we should not be permitted to adjust the distribution indefinitely in accordance with the results of one set of observations, and (6) would not hold. 'Student' indeed noticed this, for his original tables† are entitled 'Tables for estimating the probability that the mean of a unique sample of observations lie between  $-\infty$  and any given distance of the mean of the population from which the sample is drawn'. There is no particular virtue in the word 'unique' if the probability is on data  $x$ ,  $\sigma$ ,  $H$ ; the rule (1) would apply to every sample separately. But when the problem is to proceed from the sample to  $x$  uniqueness is important. If  $H$  contained information from a previous sample, this would not affect (1), since, given  $x$  and  $\sigma$ , any further information about them would tell us nothing new.

† *Biometrika*, 11, 1917, 414.

But it would affect the transition from (1) to (5), and this would be recognized in practice by combining the samples and basing the estimate on the two together. 'Student' called my attention to the vital word just after the publication of a paper of mine on the subject,<sup>†</sup> showing that he had in fact clearly noticed the necessity of the condition that the sample considered must constitute our *only* information about  $x$  and  $\sigma$ . The conditions contemplated by him are in fact completely identical with mine, and he recognized the essential point, that the usefulness of the result depends on the particular state of previous knowledge, namely, absence of knowledge.

It can be shown further that if we take (4) as giving the correct posterior probability of  $x$ , there is only one distribution of the prior probability that can lead to it, namely

$$P(dx d\sigma | H) \propto dx d\sigma / \sigma. \quad (7)$$

For the result implies that the most probable value of  $x$  is the mean, and that for two observations there is a probability  $\frac{1}{2}$  that  $x$  lies between them. But the former implies a uniform prior probability distribution for  $x$ , and the latter, by 3.7, implies the  $d\sigma/\sigma$  rule.<sup>‡</sup> Given this my argument in 3.4 follows. The irrelevance of information in the sample other than  $\bar{x}$  and  $s$  holds for all assessments of the prior probability. Hence the hypotheses made by 'Student' are completely equivalent to mine; they have merely been introduced in a different order.

Similar considerations affect Fisher's fiducial argument. Speaking of 'Student's' rule, he says:<sup>§</sup> 'It must now be noticed that  $t$  is a continuous function of the unknown parameter, the mean, together with observable values,  $\bar{x}$ ,  $s$ , and  $n$ , only. Consequently the inequality

$$t > t_1$$

is equivalent to the inequality

$$\mu < \bar{x} - st_1/\sqrt{n}$$

so that this last must be satisfied with the same probability as the first. . . . We may state the probability that  $\mu$  is less than any assigned value, or the probability that it lies between any assigned values, or, in short, its probability distribution, in the light of the sample observed.' The innocent-looking mathematical transformation, however, covers the passage from data  $x$  and  $\sigma$  to data  $\bar{x}$  and  $s$  (Fisher's  $\mu$  being my  $x$ ) which the notation used is not adequate to express. The original assessment was on data including  $\mu$ , and if these were still being used the

<sup>†</sup> *Proc. Roy. Soc. A*, 160, 1937, 325-48.

<sup>‡</sup> A proof adapted to the normal law of error is given in my paper just mentioned.

<sup>§</sup> *Ann. Eugen.* 6, 1935, 392.

probability that  $\mu$  is in a particular range is 1 if the range includes the known value and 0 if it does not. The argument therefore needs the same elaboration as was applied above to that of 'Student'. It may be noticed that in speaking of the probability distribution of  $\mu$  in the light of the sample Fisher has apparently abandoned the restriction of the meaning of probability to direct probabilities; different values of  $\mu$  are different hypotheses and he is speaking of their probabilities on the data, apparently, in precisely the same sense as I should. He does criticize the use of the prior probability in the same paper, but he appears to understand by it something quite different from what I do. My only criticism of both his argument and 'Student's' is that they omit important steps, which need considerable elaboration, and that when these are given the arguments are much longer than those got by introducing the prior probability to express previous ignorance at the start.

Fisher heads a section in his book† 'The significance of the mean of a unique sample' and proceeds: 'If  $x_1, x_2, \dots, x_n$  is a sample of  $n$  values of a variate  $x$ , and if this sample constitutes the whole of the information on the point in question, then we may test whether the mean of  $x$  differs significantly from zero by calculating the statistics. . . .' Here we have the essential point made perfectly explicit. The test is not independent of previous knowledge, as Fisher is liable to say in other places; it is to be used only where there is no relevant previous knowledge. 'No previous knowledge' and 'any conditions of previous knowledge' differ as much as 'no money' and 'any amount of money' do.

7.11. A different way of justifying the practical use of the rule without speaking of the probability of different values of  $x$  is as follows. Since  $P(dz | x, \sigma, H)$  is independent of  $x$  and  $\sigma$ , and of all previous observations, it is a chance. If we take an enormous number of samples of number  $n$ , the fraction with  $z$  between two assigned values will approximate to the integral of the law between them, by Bernoulli's theorem.

This will be true whether  $x$  and  $\sigma$  are always the same or vary from one sample to another. Then we can apparently say that actual values of  $z$  will be distributed in proportion to the integrals of  $(1+z^2)^{-1/2n}$  and regard actual samples as a selection from this population; then the probabilities of errors greater than  $\pm zs$  will be assigned in the correct ratio by the rule that the most probable sample is a fair sample. The trouble about the argument, however, is that it would hold equally well if  $x$  and  $\sigma$  were the same every time. If we proceed to say that  $x$  lies between  $\bar{x} \pm 0.75s$  in every sample of ten observations that we make, we shall be

† *Statistical Methods*, 1936, p. 125.

wrong in about 5 per cent. of the cases, irrespective of whether  $x$  is the same every time or not, or of whether we know it or not. It is suggested that we should habitually reject a suggested value of  $x$  by some such rule as this, but applying this in practice would imply that if  $x$  was known to be always the same we must accept it in 95 per cent. and reject it in 5 per cent. of the cases, which hardly seems a satisfactory state of affairs. There is no positive virtue in rejecting a hypothesis in 5 per cent. of the cases where it is true, though it may be inevitable, if we are to have any rule at all for rejecting it when it is false, that we shall sometimes reject it when it is true. In practice nobody would use the rule in this way if  $x$  was always the same; samples would always be combined. Thus, whatever may be recommended in theory, the statistician does allow for previous knowledge by the rather drastic means of restricting the range of hypotheses that he is willing to consider at all. The rule recommended would be used only when there is no previous information relevant to  $x$  and  $\sigma$ . Incidentally Bernoulli's theorem, interpreted to give an inference about what *will* happen in a large number of trials, cannot be proved from a frequency definition, and the passage to an inference in a single case, which is the usual practical problem, still needs the notion of degree of reasonable belief, which therefore has to be used twice.

Some hypothesis is needed in any case to enable us to proceed from a comparison of different sets of data on the same hypothesis to a comparison of different hypotheses on the same data; no discredit is therefore to be attached to 'Student' for making one. It cannot, however, be claimed legitimately that the argument is independent of previous knowledge. It would be valid only in the special case where there is no previous knowledge about  $x$  and  $\sigma$ , and would not be used in practice in any other. The hypothesis that, given  $H$  but no information about  $x$  and  $\sigma$  other than that provided by  $\bar{x}$  and  $s$ ,  $\bar{x}$  and  $s$  are irrelevant to  $z$  is essential to the argument. It may be accepted as reasonable, but it is none the less a hypothesis.

**7.2.** An enigmatic position in the history of the theory of probability is occupied by Karl Pearson. His best-appreciated contributions in principle are perhaps the invention of  $\chi^2$ , the introduction of the product moment formula to estimate the correlation coefficient, and the Pearson types of error law; besides of course an enormous number of applications to special subjects. I should add to these the *Grammar of Science*, which remains the outstanding general work on scientific method, and the recognition in it that the Bayes-Laplace uniform assessment of the

prior probability is not final, but can be revised to take account of previous information about the values that have occurred in the past in analogous problems. The anomalous feature of his work is that though he always maintained the principle of inverse probability, and made this important advance, he seldom used it in actual applications, and usually presented his results in a form that appears to identify a probability with a frequency. In particular his numerous tables of chances are mostly entitled frequencies. In determining the parameters of laws of his own types from observations he did not use inverse probability, and when Fisher introduced maximum likelihood, which is practically indistinguishable from inverse probability in estimation problems, Pearson continued to use the method of moments. A possible reason for this that many would appreciate is that complete tables for fitting by moments were already available, and that the fitting of a law with four adjustable parameters by maximum likelihood is not a matter to be undertaken lightly when sufficient statistics do not exist. But Pearson in his very last paper maintained that the method of moments was not merely easier than maximum likelihood, but actually gave a better result. He also never seems to have seen the full importance of  $\chi^2$  itself. When the data are observed numbers, he showed that the probability of the numbers, given a law, is proportional to  $\exp(-\frac{1}{2}\chi^2)$  with a third-order error. Thus the equivalence of maximum likelihood and minimum  $\chi^2$  was Pearson's result, and the close equivalence of maximum likelihood and inverse probability in estimation problems is so easy to show that it is remarkable that Pearson overlooked it. Most of the labour of computing the likelihood is avoided if  $\chi^2$  is used instead, though there are complications when some of the expectations are very small; but even these are avoided by the treatment of 4.2. Fisher repeatedly drew attention to the relation between maximum likelihood and minimum  $\chi^2$ , but Pearson never accepted the consequence that if he used the latter he would have had a convenient method, more accurate than the method of moments, and justified by principles that he himself had stated repeatedly.

In practice Pearson used  $\chi^2$  only as a significance test. His method, if there were  $n$  groups of observations, was to compute the complete  $\chi^2$  for the data, in comparison with the law being tested. If  $m$  parameters had been found from the data, he would form the integral

$$P(\chi^2) = \int_{\chi^2}^{\infty} \chi^{n-m-1} e^{-1/2\chi^2} d\chi / \int_0^{\infty} \chi^{n-m-1} e^{-1/2\chi^2} d\chi,$$



which is the probability, given a law, that the  $\chi^2$  formed from  $n-m$  random variations in comparison with their standard errors would exceed the observed value. (In his earlier use of  $\chi^2$  he allowed only for one adjustable parameter, the whole number of observations; the need to allow for all was pointed out by Fisher† and emphasized by Yule.‡) If  $P$  was less than some standard value, say 0.05 or 0.01, the law considered was rejected. Now it is with regard to this use of  $P$  that I differ from all the present statistical schools, and detailed attention to what it means is needed. The fundamental idea, and one that I should naturally accept, is that a law should not be accepted on data that themselves show large departures from its predictions. But this requires a quantitative criterion of what is to be considered a large departure. The probability of getting the whole of an actual set of observations, given the law, is ridiculously small. Thus for frequencies 2.74 (6) shows that the probability of getting the observed numbers, in any order, decreases with the number of observations like  $(2\pi N)^{-1/2(p-1)}$  for  $\chi^2 = 0$  and like  $(2\pi Ne)^{-1/2(p-1)}$  for  $\chi^2 = p-1$ , the latter being near the expected value of  $\chi^2$ . The probability of getting them in their actual order requires division by  $N!$ . If mere improbability of the observations, given the hypothesis, was the criterion, any hypothesis whatever would be rejected. Everybody rejects the conclusion, but this can mean only that improbability of the observations, given the hypothesis, is not the criterion, and some other must be provided. The principle of inverse probability does this at once, because it contains an adjustable factor common to all hypotheses, and the small factors in the likelihood simply combine with this and cancel when hypotheses are compared. But without it some other criterion is still necessary, or any alternative hypothesis would be immediately rejected also. Now the  $P$  integral does provide one. The constant small factor is rejected, for no apparent reason when inverse probability is not used, and the probability of the observations is replaced by that of  $\chi^2$  alone, one particular function of them. Then the probability of getting the same or a larger value of  $\chi^2$  by accident, given the hypothesis, is computed by integration to give  $P$ . If  $\chi^2$  is equal to its expectation supposing the hypothesis true,  $P$  is about 0.5. If  $\chi^2$  exceeds its expectation substantially, we can say that the value would have been unlikely to occur had the law been true, and shall naturally suspect that the law is false. So much is clear enough. If  $P$  is small, that means that there have been unexpectedly large departures from prediction. But why

† *J. R. Stat. Soc.* 85, 1922, 87-94.

‡ *Ibid.*, pp. 95-106.

should these be stated in terms of  $P$ ? The latter gives the probability of departures, measured in a particular way, equal to or greater than the observed set, and the contribution from the actual value is nearly always negligible. *What the use of  $P$  implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred.* This seems a remarkable procedure. On the face of it the fact that such results have not occurred might more reasonably be taken as evidence for the law, not against it. The same applies to all the current significance tests based on  $P$  integrals.†

The use of the integral goes back to Chauvenet's criterion for rejecting observations. This proceeded as follows. Let  $P(m)$  be the chance on the normal law of an error greater than  $m\sigma$ . Then the chance that all of  $n$  errors will be less than  $m\sigma$  is  $\{1 - P(m)\}^n$ , and the chance that there will be at least one greater than  $m\sigma$  is  $1 - \{1 - P(m)\}^n$ . The first estimate of the true value and standard error were used to find the chance that there would be at least one residual larger than the largest actually found. If this was greater than  $\frac{1}{2}$  the observation was rejected, and a mean and a standard error were found from the rest and the process repeated until none were rejected. Thus on this method there would be an even chance of rejecting the extreme observation even if the normal law was true. If such a rule was used now the limit would probably be drawn at a larger value, but the principle remains, that an observation that might be normal is rejected because other observations not predicted by the law have not occurred. Something might be said for rejecting the extreme observation if the law gave a small chance of a residual exceeding the second largest; then indeed something not predicted by the law might be said to have occurred, but to apply such a rule to the largest observation is wrong in principle. (Even if the normal law does not hold, rejection of observations and treating the rest as derived from the normal law is not the best method, and may give a spurious accuracy; but the question here concerns the decision as to whether the normal law applies to all the  $n$  observations.)

It must be said that the method fulfils a practical need; but there was no need for the paradoxical use of  $P$ . The need arose from the fact that in estimating new parameters the current methods of estimation ordinarily gave results different from zero, but it was habitually found

† On the other hand, Yates (*J. R. Stat. Soc., Suppl.* 1, 1934, 217-35) recommends, in testing whether a small frequency  $n_r$  is consistent with expectation, that  $\chi^2$  should be calculated as if this frequency was  $n_r + \frac{1}{2}$  instead of  $n_r$ , and thereby makes the actual value contribute largely to  $P$ . This is also recommended by Fisher (*Statistical Methods*, p. 98). It only remains for them to agree that nothing but the actual value is relevant.

that those up to about twice the standard error tended to diminish when the observations became more numerous or accurate, which was what would be expected if the differences represented only random error, but not what would be expected if they were estimates of a relevant new parameter. But this could be dealt with in a rough empirical way by taking twice the standard error as a criterion for possible genuineness and three times the standard error for definite acceptance. This would rest on a valid inductive inference from analogous cases, though not necessarily the best one. Now this would mean that the former limit would be drawn where the joint probability of the observations is  $e^{-2}$  of the value for the most probable result, supposing no difference present, and the latter at  $e^{-4.5}$ . This would depend on the probability of the actual observations and thus on the ordinate of the direct probability distribution, not on the integral. The ordinate does depend on the hypothesis and the observed value, and nothing else. Further, since nearly all the more accurate tests introduced since have depended on the use of distributions that are nearly normal in the range that matters, there would be a natural extension in each case, namely to draw the two lines where the ordinates are  $e^{-2}$  and  $e^{-4.5}$  times those at the maximum. The practical difference would not be great, because in the normal distribution, for instance, for  $x$  large and positive,

$$\frac{1}{\sqrt{(2\pi)\sigma}} \int_x^\infty \exp\left(-\frac{x^2}{2\sigma^2}\right) dx \sim \sqrt{\left(\frac{2}{\pi}\right)\frac{\sigma}{x}} \exp\left(-\frac{x^2}{2\sigma^2}\right)$$

and the exponential factor varies much more rapidly than  $x$ . The use of a standard value for the ordinate rather than  $P$  would give practically the same decisions in all such cases. Its choice, however, would rest on inductive evidence, which could be stated; there would be no need for the apparently arbitrary choice of fixed limits for  $P$ , or for the paradox in the use of  $P$  at all.

Some feeling for the ordinate seems to lie behind the remarks (see p. 281) of Yule and Kendall and Fisher on the subject of suspiciously small  $\chi^2$  and  $P$  very near 1. It is hard to understand these if  $P$  is taken as the sole criterion, but they become comprehensible at once if the ordinate is taken as the criterion;  $P$  very near 1 does correspond to a small ordinate.

**7.21.** It should be said that several of the  $P$  integrals have a definite place in the present theory, in problems of pure estimation. For the normal law with a known standard error, or for those sampling problems that reduce to it, the total area of the tail represents the probability,

given the data, that the estimated difference has the right sign—provided that there is no question whether the difference is zero. (If some previous suggestion of a specific value of a parameter is to be considered at all, it must be disposed of by a significance test before any question of estimating any other value arises. Then, strictly speaking, if the adjustable parameter is supported by the data the test gives its posterior probability as a by-product.) Similarly, the  $t$  rule gives the complete posterior probability distribution of a quantity to be estimated from the data, provided again that there is no doubt initially about its relevance; and the integral gives the probability that it is more or less than some assigned value. The  $z$  rule also gives the probability distribution of the scatter of a new set of observations or of means of observations, given an existing set. These are all problems of pure estimation. But their use as significance tests covers a looseness of statement of what question is being asked. They give the correct answer if the question is: If there is nothing to require consideration of some special values of the parameter, what is the probability distribution of that parameter given the observations? But the question that concerns us in significance tests is: If some special value has to be excluded before we can assert any other value, what is the best rule, on the data available, for deciding whether to retain it or adopt a new one? The former is what I call a problem of estimation, the latter of significance. Some feeling of discomfort seems to attach itself to the assertion of the special value as *right*, since it may be slightly wrong but not sufficiently to be revealed by a test on the data available; but no significance test asserts it as certainly right. We are aiming at the best way of progress, not at the unattainable ideal of immediate certainty. What happens if the null hypothesis is retained after a significance test is that the maximum likelihood solution or a solution given by some other method of estimation is rejected. The question is, When we do this, do we expect thereby to get more or less correct inferences than if we followed the rule of keeping the estimation solution regardless of any question of significance? I maintain that the only possible answer is that we expect to get more. The difference as estimated is interpreted as random error and irrelevant to future observations. In the last resort, if this interpretation is rejected, there is no escape from the admission that a new parameter may be needed for every observation, and then all combination of observations is meaningless, and the only valid presentation of data is a mere catalogue without any summaries at all.

If any concession is to be made to the opinion that a new parameter

rejected by a significance test is probably not zero, it can be only that it is considerably less than the standard error given by the test; but there is no way of stating this sufficiently precisely to be of any use.

The use of the  $P$  integral in significance tests, however, merely expresses a feeling that some standard is required. In itself it is fallacious because it rejects a hypothesis on account of observations that have not occurred; its only justification is that it gives some sort of a standard which works reasonably well in practice, but there is not the slightest reason to suppose that it gives the best standard. Fisher writes,<sup>†</sup> speaking of the normal law: 'The value for which  $P = 0.05$ , or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not. Deviations exceeding twice the standard error are thus formally regarded as significant. Using this criterion we should be led to follow up a false indication only once in 22 trials, even if the statistics were the only guide available. Small effects will still escape notice if the data are insufficiently numerous to bring them out, but no lowering of the standard of significance would meet this difficulty.' *Convenient* is Fisher's word; there is no claim that the criterion is the best. But the idea that the best limit can be drawn at some unique value of  $P$  has somehow crept into the literature, without apparently the slightest attempt at a justification or any ground for saying what the best value is.

The distinction between problems of estimation and significance arises in biological applications, though I have naturally tended to speak mainly of physical ones. Suppose that a Mendelian finds in a breeding experiment 459 members of one type, 137 of the other. The expectations on the basis of a 3:1 ratio would be 447 and 149. The difference would be declared not significant by any test. But the attitude that refuses to attach any meaning to the statement that the simple rule is right must apparently say that if any predictions are to be made from the observations the best that can be done is to make them on the basis of the ratio 459/137, with allowance for the uncertainty of sampling. I say that the best is to use the 3/1 rule, considering no uncertainty beyond the sampling errors of the new experiments. In fact the latter is what a geneticist would do. The observed result would be recorded and might possibly be reconsidered at a later stage if there was some question of differences of viability after many more observations had accumulated; but meanwhile it would be regarded as confirmation of the theoretical value. This is a problem of what I call significance.

<sup>†</sup> *Statistical Methods*, p. 46.

But what are called significance tests in agricultural experiments seem to me to be very largely problems of pure estimation. When a set of varieties of a plant are tested for productiveness, or when various treatments are tested, it does not appear to me that the question of presence or absence of differences comes into consideration at all. It is already known that varieties habitually differ and that treatments have different effects, and the problem is to decide which is the best; that is, to put the various members, as far as possible, in their correct order. The design of the experiment is such that the order of magnitude of the uncertainty of the result can be predicted from similar experiments in the past, and especially from uniformity trials, and has been chosen so that any differences large enough to be interesting would be expected to be revealed on analysis. The experimenter has already a very good idea of how large a difference needs to be before it can be considered to be of practical importance; the design is made so that the uncertainty will not mask such differences. But then the  $P$  integral found from the difference between the mean yields of two varieties gives correctly the probability on the data that the estimates are in the wrong order, which is what is required. If the probability that they are misplaced is under 0.05 we may fairly trust the decision. It is hardly correct in such a case to say that previous information is not used; on the contrary, previous information relevant to the orders of magnitude to be compared has determined the whole design of the experiment. What is not used is previous information about the differences between the actual effects sought, usually for the very adequate reason that there is none; and about the error likely to arise in the particular experiment, which is only an order of magnitude and by the results found several times in this book can be treated as previous ignorance as soon as we have directly relevant information. If there are any genuine questions of significance in agricultural experiments it seems to me that they must concern only the higher interactions.

7.22. A further problem that arises in the use of any test that simply rejects a hypothesis without at the same time considering possible alternatives is that admirably stated by the Cheshire Cat in the quotation at the head of Chapter V. Is it of the slightest use to reject a hypothesis until we have some idea of what to put in its place? If there is no clearly stated alternative, and the null hypothesis is rejected, we are simply left without any rule at all, whereas the null hypothesis, though not satisfactory, may at any rate show some sort of correspondence with the facts. It may for instance represent 90 per cent. of

the variation and to that extent may be of considerable use in prediction, even though the remaining 10 per cent. may be larger than we should expect if it was strictly true. Consider, for instance, the history of the law of gravitation. Newton first derived it from Kepler's laws and a comparison of the accelerations of the moon and of a body falling freely at the earth's surface. Extending it to take account of the mutual attractions of the planets and of the perturbations of the moon by the sun, he got the periods and orders of magnitude of the principal perturbations. But he did not explain the long inequality of Jupiter and Saturn, with a period of 880 years, which gives displacements in longitude of  $1196''$  and  $2908''$  of arc for the two planets,<sup>†</sup> and was only explained by Laplace a century later. The theory of the moon has been taken only in the present century, by E. W. Brown, to a stage where the outstanding errors can be said to be within the errors of observation; and even now the theory involves the empirical secular acceleration of the mean motion, attributable to tidal friction, a periodic empirical term with an amplitude of  $10.7''$  and a period of seventy years, and some curious short-period fluctuations that are not satisfactorily explained. In fact agreement with Newton's law was not given by the data used to establish it, because these data included the main inequalities of the moon; it was not given during his lifetime, because the data included the long inequality of Jupiter and Saturn; and when Einstein's modification was adopted the agreement of observation with Newton's law was 300 times as good as Newton ever knew. Even the latter appears at present as powerless as Newton's to explain the long empirical term in the moon's longitude and the secular motion of the node of Venus. There has not been a single date in the history of the law of gravitation when a modern significance test would not have rejected all laws and left us with no law. Nevertheless the law did lead to improvement for centuries, and it was only when an alternative was sufficiently precisely stated to make verifiable predictions that Newton's law could be dropped—except of course in the cases where it is still a valid approximation to Einstein's, which happen to be most of the cases. The test required, in fact, is not whether the null hypothesis is altogether satisfactory, but whether any suggested alternative is likely to give an improvement in representing future data. If the null hypothesis is not altogether satisfactory we can still point to the apparent discrepancies as possibly needing further attention, and attention to their amount

<sup>†</sup> I am indebted for the values to Mr. D. H. Sadler; they are from G. W. Hill, *Astronomical Papers of the American Ephemeris*, vols. iv and vii.

gives an indication of the general magnitude of the errors likely to arise if it is used; and that is the best we can do.

**7.23.** The original use of  $\chi^2$  involves a further difficulty, which could occur also in using Fisher's  $z$ , which is the extension of  $\chi^2$  to take account of the uncertainty of the standard error. If we have a set of frequencies,  $n-m$  of which could be altered without producing an inconsistency with the marginal totals of a contingency table, their variations could be interpreted as due to  $n-m$  possible new functions in a law of chance, which would then give  $\chi^2 = 0$ ; or they could be due to a failure of independence, a tendency of observations to occur in bunches increasing  $\chi^2$  systematically without there necessarily being any departure from proportionality in the chances. We have seen the importance of this in relation to the annual periodicity of earthquakes. Similarly, when the data are measures they can be divided into groups and means taken for the groups. The variation of the group means can be compared with the variations in the groups to give a value of  $z$ . But this would be increased either by a new function affecting the measures or by a failure of independence of the errors, which need not be expressible by a definite function. The simple use of  $\chi^2$  or of  $z$  would not distinguish between these; each new function or a failure of independence would give an increase, which might lead to the rejection of the null hypothesis, but we shall still have nothing to put in its place until we have tested the various alternatives. What is perhaps even more serious is that with a large number of groups the random variation of  $\chi^2$  on the null hypothesis is considerable, and a systematic variation that would be detected at once if tested directly may pass as random through being mixed up with the random error due simply to the arbitrary method of grouping (cf. 2.76, p. 91). Fisher of course has attended to this point very fully, though some of his enthusiastic admirers seem to have still overlooked it. Both with  $\chi^2$  and  $z$  it is desirable to separate the possible variation into parts when the magnitude of one gives little or no information about what is to be expected of another, and to test each part separately. The additive property of  $\chi^2$  makes it easily adaptable for this purpose. Each component of variation makes its separate contribution to  $\chi^2$ , and  $\exp(-\frac{1}{2}\chi^2)$  separates into factors, so that the contributions are mutually irrelevant. It is for this reason that  $\chi^2$  and  $t^2$  have appeared explicitly in my tests where several new parameters are associated. The  $\chi^2$  here is not the complete  $\chi^2$ , but the contribution for the possible component variations directly under consideration. Whether the random variation is more



or less than its expectation (so long as it is random) is irrelevant to the test.

7.3. The treatment of expectations is another peculiar feature of Pearson's work. The choice of a set of functions of the observations, and equating them to the expectations given the law under consideration, is often a convenient way of estimating the parameters. Pearson used it habitually in the method of moments and in other work. It is not necessarily the best method, but it is liable to be the easiest. But it is often very hard to follow in Pearson's presentations and in those of some of his followers. It is indeed very difficult on occasion to say whether in a particular passage Pearson is speaking of a function of the observations or the expectation that it may be an estimate of. When he speaks of a 'mean' he sometimes intends the mean of the observations, sometimes the expectation of one observation given the law, and the complications become greater for higher moments. The transition from the function of the observations to the corresponding expectation involves a change of data, which is passed over without mention even when the use of inverse probability may be recommended a few pages later.

7.4. The general agreement between Professor R. A. Fisher and myself has been indicated already in many places. The apparent differences have been much exaggerated owing to a rather unfortunate discussion some years ago, which was full of misunderstandings on both sides. Fisher thought that a prior probability based on ignorance was meant to be a statement of a known frequency, whereas it was meant merely to be a formal way of stating that ignorance, and I had been insisting for several years that no probability is simply a frequency. I thought that he was attacking the 'Student' rule, of which my result for the general least squares problem was an extension; at the time, to my regret, I had not read 'Student's' papers and it was not till considerably later that I saw the intimate relation between his methods and mine. This discussion no longer, in my opinion, needs any attention. My main disagreement with Fisher concerns the hypothetical infinite population, which is a superfluous postulate since it does not avoid the need to estimate the chance in some other way, and the properties of chance have still to be assumed since there is no way of proving them. Another is that, as in the fiducial argument, an inadequate notation enables him, like 'Student', to pass over a number of really difficult steps without stating what hypotheses are involved in them. The third is the use of the  $P$  integral, but Fisher's alertness for possible dangers is so great

that he has anticipated all the chief ones. I have in fact been struck repeatedly in my own work, after being led on general principles to a solution of a problem, to find that Fisher had already grasped the essentials by some brilliant piece of common sense, and that his results would be either identical with mine or would differ only in cases where we should both be very doubtful. As a matter of fact I have applied my significance tests to numerous applications that have also been worked out by Fisher's, and have not yet found a disagreement in the actual decisions reached. The advantage of my treatment, I should say, is that it shows the relation of these methods among themselves, and to general principles concerning the possibility of inference, whereas in Fisher's they apparently involve independent postulates. In relation to some special points, my methods would say rather more for Fisher's than he has himself claimed. Thus he claims for maximum likelihood only that it gives a systematic error of order less than  $n^{-1/2}$  in the ordinary cases where the standard error is itself of order  $n^{-1/2}$ . Inverse probability makes the systematic error of order  $n^{-1}$ . He shows also by a limiting argument that statistics given by the likelihood lead to estimates of the population parameters at least as accurate as those given by any other statistics, when the number of observations is large. Inverse probability gives the result immediately without restriction on the number of observations. The fiducial argument really involves hypotheses equivalent to the use of inverse probability, but the introduction of maximum likelihood appears in most cases to be an independent postulate in Fisher's treatment. In mine it is a simple consequence of general principles. The trouble about taking maximum likelihood as a primitive postulate, however, is that it would make significance tests impossible, just as the uniform prior probability would. The maximum likelihood solution would always be accepted and therefore the simple law rejected. In actual application, however, Fisher uses a significance test based on  $P$  and avoids the need to reject the simple law whether it is true or not; thus he gets common-sense results though at the cost of some sacrifice of consistency. The point may be illustrated by a remark of W. G. Emmett† to the effect that if an estimated difference  $t$  is less than the adopted limit, it affords no ground for supposing the true difference to be 0 rather than  $2t$ . If we adopted maximum likelihood or the uniform prior probability in general there would be no escape from Emmett's conclusion; but no practical statistician would accept it. Any significance test whatever involves the recognition that there is

† *B. J. Psych.* 26, 1936, 362-87.

something special about the value 0, implying that the simple law may possibly be true; and this contradicts the principle that the maximum likelihood estimate, or any unbiased estimate, is always the best.

Fisher has already introduced the useful word 'fiducial' for limits, in estimation problems, such that there may be on the data a specified probability that the true value lies between them. But it seems to be supposed that 'fiducial' and 'significant' mean the same thing, which is not the case.

He has often argued for making a decision rest on the observations immediately under consideration and not on any previous evidence. This appears to contradict the view that I have developed, that the best inference must rest on the whole of the relevant evidence if we are to be consistent. The difference is not so great as it appears at first sight, however. I find that vaguely recorded evidence is just as well ignored, and precisely recorded evidence may require a significance test to establish its relevance. He also avoids the tendency of the human mind to remember what it wants to believe and forget the rest, unless it is written down at the time. With such exceptions as these, with respect to which we should concur, Fisher seems to be as willing in practice to combine data as I am. In fact, in spite of his occasional denunciations of inverse probability I think that he has succeeded better in making use of what it really says than many of its professed users have.

7.5. E. S. Pearson and J. Neyman have given an extended analysis of significance tests. In any test, if we are to have a rule for detecting the falsehood of a law, we must expect to make a certain number of mistakes owing to occasional large random errors. If we habitually use a 5 per cent. *P* limit, the null hypothesis will in the ordinary course of events be rejected in about 5 per cent. of the cases where it is true. As it will often be false, if we choose such a limit the number of such mistakes will be less than 5 per cent. of the whole number of cases. It is in this sense that Fisher speaks of 'exact tests of significance'. Pearson and Neyman, however, go further. This type of mistake is called an error of the first kind. But it is also possible that a new parameter may be required and that, owing either to its smallness or to the random error having the opposite sign, the estimate is within the range of acceptance of the null hypothesis; this they call an error of the second kind, that of accepting the null hypothesis when it is false. They have given extensive discussions of the chances of such errors of the second kind,

tabulating their risks for different possible values of the new parameter.† I do not think that they have stated the question correctly, however, though this attention to errors of the second kind bears some resemblance to the principle that I have used here, that there is no point in rejecting the null hypothesis until there is something to put in its place. Their method gives a statement of the alternative. But in a practical case the alternative will either involve an adjustable parameter or will be as definitely stated as the null hypothesis. For instance, the laws of gravitation and light of Newton and Einstein involve the same number of adjustable parameters, the constant of gravitation and the velocity of light appearing in both. Now Pearson and Neyman proceed by working out the above risks for different values of the new parameter, and call the result the power function of the test, the test itself being in terms of the  $P$  integral. But if the actual value is unknown the value of the power function is also unknown; the total risk of errors of the second kind must be compounded of the power functions over the possible values, with regard to their risk of occurrence. On the other hand, if the alternative value is precisely stated I doubt whether anybody would use the  $P$  integral at all; if we must choose between two definitely stated alternatives we should naturally take the one that gives the larger likelihood, even though each may be within the range of acceptance of the other. To lay down an order of test in terms of the integral in such a case would be very liable to lead to accepting the first value suggested even though the second may agree better with the observations.

It may, however, be interesting to see what would happen if the new parameter is needed as often as not, and if the values when it is needed are uniformly distributed over the possible range. Then the frequencies in the world would be proportional to my assessment of the prior probability. Suppose, then, that the problem is, not knowing in any particular case whether the parameter is 0 or not, to identify the cases so as to have a minimum total number of mistakes of both kinds. Using the notation of 5.0, the chance of  $q$  being true and of  $a$  being in a range  $da$  is  $P(q da | H)$ . That of  $q'$ , with  $\alpha$  in a range  $d\alpha$ , and of  $a$  being in the range  $da$ , is  $P(q' d\alpha da | H)$ . If, then, we assign an  $a_c$  and assert  $q$  when  $|a| < a_c$  and  $q'$  when  $|a| > a_c$ , and sampling is random, the expectation of the total fraction of mistakes will be

$$2 \int_{a_c}^{\infty} P(q da | H) + 2 \int_0^{a_c} \int P(q' d\alpha da | H), \quad (1)$$

† *Univ. Coll. Lond., Stat. Res. Memos.* 2, 1938, 25–57, and earlier papers.

the second integral being over the range of  $\alpha$ . Thus the second integral is  $2 \int_0^{\alpha_c} P(q' da | H)$ . Now if  $\alpha_c$  is chosen to make the total a minimum, we must have for small variations about  $\alpha_c$

$$P(q da | H) = P(q' da | H). \quad (2)$$

But these are respectively equal to

$$P(da | H)P(q | \alpha_c H) \quad \text{and} \quad P(da | H)P(q' | \alpha_c H);$$

whence

$$P(q | \alpha_c H) = P(q' | \alpha_c H). \quad (3)$$

But this is the relation that defines the critical value. Hence, with world-frequencies in proportion to the prior probability used to express ignorance, the total number of mistakes will be made a minimum if the line is drawn at the critical value that makes  $K = 1$ .

Now I do not say that this proportionality holds; all that I should say myself is that at the outset we should expect to make a minimum number of mistakes in this way, but that accumulation of information may lead to a revision of the prior probabilities for further use and the critical value may be correspondingly somewhat altered. But whatever the frequency law may be, we notice that it is the values of  $\alpha$  near  $\alpha_c$  and therefore, in the cases needing discussion, the small values, that contribute most of the second term in (1). Revision would therefore alter (3) in the ratio of the numbers of the cases of  $\alpha = 0$  and of small values of  $\alpha$ , and therefore  $K$  would be altered by a factor independent of the number of observations. We should therefore get the best result, with any distribution of  $\alpha$ , by some form that makes the ratio of the critical value to the standard error increase with  $n$ . It appears then that whatever the distribution may be, the use of a fixed  $P$  limit cannot be the one that will make the smallest number of mistakes. The absolute best is of course unknown since we do not know the distribution in question except so far as we can infer it from similar cases.

**7.51.** This procedure has some interest in relation to 'giving a theory every chance'. There are cases where there is no positive evidence for a new parameter, but important consequences might follow if it was not zero, and we must remember that  $K > 1$  does not prove that it is zero, but merely that it is more likely to be zero than not. Then it is worth while to examine the alternative  $q'$  further and see what limits can be set to the new parameter, and thence to the consequences of introducing it. This occurred in the discussion of the viscosity of the earth. The new parameter here would be the rate of distortion per unit stress when the stress is maintained indefinitely long; if it is zero the viscosity is

infinite and the strength is finite. There was no positive evidence that the parameter is not zero, but if it was the way might be open to large distortions under forces acting for a long enough time. It was therefore desirable to consider what limits could be assigned to the new parameter from evidence actually available, and to see whether they would permit the amounts of distortion that were claimed. Here the use of deduction as an approximation would not permit the discussion of  $q'$  at all, but on recognizing that it is only an approximation we are free to continue to consider  $q'$  and fix limits to its consequences. It was actually found† that the largest admissible value of the new parameter, that is, the smallest possible viscosity, led to insufficient distortion under any force suggested. This is a case where a hypothesis, that of ultimate indefinitely large distortion, is disposed of not only by the lack of positive evidence for the new parameter needed to make it possible at all, but also by the fact that even on choosing the new parameter to be as favourable as possible to it, consistently with other evidence, the result is still contradicted.

**7.6.** The analysis of this chapter is relevant to the standard presentations of statistical mechanics, those of Boltzmann and Gibbs. The original derivation of the distribution of velocities, that of Maxwell, proceeded by supposing, first, that the probability of a given resultant velocity is a function of that velocity alone; secondly, that those for the three components separately are independent. From these hypotheses Maxwell's law follows. Boltzmann attempted to go more into detail by considering the probable effects of collisions, and appeared to show that a function  $H$ , representing the departure from a Maxwellian state, would diminish. An objection to Maxwell's treatment was that he assumed independence of the components. But he claimed only to consider the steady state, where this might possibly hold. Boltzmann, however, considered departures from the steady state, and assumed irrelevance between the positions and velocities of neighbouring molecules. This is plainly illegitimate if the density is not uniform or if the velocity varies systematically between regions. The presence of one molecule in a region affords ground for supposing that the region is one of high density and therefore gives an excess probability that there will be another near to it. A velocity of a molecule implies an excess probability that a neighbour has one in a similar direction; in each case supposing that any original departures from homogeneity have not had

† Jeffreys, *The Earth*, 1929, pp. 304–5.

time to be smoothed out. Thus Boltzmann's treatment is definitely worse than Maxwell's, in spite of its greater complexity. Maxwell applied the hypothesis of independence only to the case where it might be true; Boltzmann applied it to cases where it quite certainly contradicts the premisses. His argument affords no ground whatever for supposing that a system will approach a Maxwellian state, because it is only when the final state has been reached that the hypotheses can possibly be right. This criticism of the Boltzmann method would be appreciated by any statistician that understands a correlation coefficient.

In the treatment of Gibbs no attempt is made to treat the individual system; instead, an ensemble of an infinite number is set up and conclusions are drawn as averages over the ensemble. But there is no guarantee at all that an average has any relevance to a single system. It might, for instance, be merely the mean of two peaks and itself correspond to no individual case at all. What is done is to consider the state of a system by regarding the  $n$  coordinates and  $n$  momenta as plotted in space of  $2n$  dimensions. Then the values at any instant determine the rates of change, by the equations of dynamics, and we can consider how the volume of a small region (corresponding to a range of different systems) will vary if each point in it moves at the rate so specified. Liouville's theorem shows that it will not vary. By some process that is recognized as obscure this is made to lead to the conclusion that the density in this phase space is uniform. Thus Jeans† appeals to experiment to say that if a property is found to hold in general for systems that have been left to themselves for a long time, that must mean either that the representative points crowd into the regions where that property holds, which is forbidden by Liouville's theorem; or that the property is true for the whole of the space, and therefore, apparently, the distribution of density does not matter and may as well be taken uniform. But there is no theoretical reason to show that there should be any such properties. Fowler‡ gives a similar argument, including the statement 'that such a  $W$  really exists is largely a pious hope'. What can be done by these methods is at the most to obtain relations between properties, assuming that such relations exist; they give no explanation of why they should exist. This can be done only by considering the individual system and showing that certain properties would be expected to hold for any individual system. Any sort of averaging is definitely dangerous.

The fundamental fact appears to be that we do not in general know

† *Dynamical Theory of Gases*, 1921, p. 73.

‡ *Statistical Mechanics*, 1929, p. 12.

the initial state of the system sufficiently accurately to predict even one collision. Though the equations of classical mechanics would ordinarily lead to a unique solution if the initial state was known exactly, and we had enough time for the computation, a trifling uncertainty in the velocity of one molecule would affect the identity of the first struck by it, and this would lead to differences afterwards that would ultimately affect the entire system. It is this uncertainty that requires the introduction of probability at all. For a system with exactly known initial conditions there would be a unique trajectory in phase space (classical mechanics of course being assumed). But for the actual system we have a set of possible trajectories with different probabilities forming a continuous set. On account of the collisions, even if these differ only slightly originally, they will quickly become widely scattered. The essential point is not so much that the volume of an element in the phase space remains the same as that its shape is distorted continuously between every pair of collisions, and it is broken up and displaced bodily at every collision. The result is that if we fix attention on a given element of the phase space, the chance that the system will be within it after a long time is made up of components from the probabilities of all the possible initial states. The tendency of this averaging is to make the probability density after a long time uniform, subject to the condition that the only admissible states are those with the same invariant properties as the original state—such as energy, for all conservative systems, and linear and angular momentum, for free systems. The density in phase space thus acquires a definite meaning as a true probability, arising ultimately from the fact that we do not know the initial state accurately. It leads to inferences about, for instance, the probability that there will be a given fraction of the momenta in one direction between stated limits, and hence to definite predictions about statistical properties such as pressure and density for every individual system. Thus the theory does give what is wanted, a prediction about the ultimate state of the individual system and made with practical certainty.† It is in no other sense that the relations found can be considered as physical laws or the quantities in them as physical magnitudes.

The general principles of this kind of averaging are known as ergodic theory and have been extensively studied, especially by French and Russian authors.‡

† *Proc. Roy. Soc. A*, **160**, 1937, 337–47.

‡ Cf. M. Fréchet, Borel's *Traité du calcul des probabilités*, t. 1, fasc. 3, 1938; H. and B. S. Jeffreys, *Methods of Mathematical Physics*, 1946, 148–52.



## VIII

### GENERAL QUESTIONS

'But you see, I can believe a thing without understanding it. It's all a matter of training.'

DOROTHY L. SAYERS, *Have His Carcase*.

8.0. Most of the present books on statistics, and of the longer papers in journals, include a careful disclaimer that the authors propose to use inverse probability, and emphasize its lack of logical foundation, which is supposed to have been repeatedly pointed out. In fact the continued mention of a principle that everybody is completely convinced is nonsense recalls the saying of the Queen in *Hamlet*: 'The lady doth protest too much, methinks.' Unfortunately some people that have examined the question have not been so convinced, and they include such first-rate logicians as W. E. Johnson, C. D. Broad, and F. P. Ramsey. The objectors, however, mostly seem to understand by the principle something so nonsensical that it hardly seems worth attention, namely that the prior probability is intended to be a known frequency. This statement has been repeated by Kendall† since the first edition of this book. *The essence of the present theory is that no probability, direct, prior, or posterior, is simply a frequency.* The fundamental idea is that of a reasonable degree of belief, which satisfies certain rules of consistency and can in consequence of these rules be formally expressed by numbers by means of the addition rule, which in itself is a convention. In many cases the numerical assessment is the same as that of a corresponding frequency, but that does not say that the probability and the frequency are the same thing even in those cases. The fact that physicists describe an atmospheric pressure as 759 millimetres does not make a pressure into a length (and meteorologists now give the pressure in terms of the millibar, which really is a unit of pressure). A number of choices of units so that certain constants of proportionality would have measure unity, and then the identification of the constants with the number unity, led to the amazing conclusion that the ratio of the electrostatic and electromagnetic units of charge, which are quantities of the same kind, is the velocity of light; and instead of seeing that this was a *reductio ad absurdum* several generations of physicists tried to justify it. There are signs now that the fact is appreciated. The equations of heat conduction and diffusion have the same form, but that does not make heat a vapour. The notion of a reasonable degree

† *The Advanced Theory of Statistics*, 1, 178.

of belief must be brought in before we can speak of a probability; and even those writers that do not mention it at the beginning have to use it at the end before any application can be made of the results—or else avoid the question by allowing the person advised to supply it himself, which he does in practice without the slightest difficulty. Even if the prior probability is *based on* a known frequency, as it is in some cases, reasonable degree of belief is needed before any use can be made of it. It is not *identical with* the frequency.

The kind of case where a prior probability may be based on a known frequency is the following. Suppose (a) we deliberately make up 10,001 classes of 10,000 balls each, such that one contains 10,000 white ones, the next 9,999 white and 1 black, and so on. We select one of these at random and extract a sample of 30, 20 of which are found to be white and 10 black. By the condition of randomness the chance of selecting any class for sampling is the same, and the prior probability for its composition follows Laplace's rule. We infer that in the class sampled about  $\frac{2}{3}$  are probably white and the rest black, the probabilities for other ratios being distributed according to a definite rule. But suppose (b) that classes of 10,000 were chosen at random from a class of number  $10^{10}$ , about the composition of which we had no previous information, and that we again sampled one of them and found 20 white and 10 black balls. Again the prior probability follows Laplace's rule, but for a different reason. The posterior probabilities for the class sampled are the same in both cases. Case (b) is the one that usually concerns us, but the analysis is quite capable of dealing with (a), in which the prior probability is based on a known frequency. It may be pointed out that if we take a sample from a second class there will be a considerable difference in the results in the two cases. For in (a) the probability that the composition will have any particular value is almost what it was before; the only difference is that since one class, whose ratio was probably near 2:1, has been excluded, the probability that the second class will yield a sample with a composition in this neighbourhood is a shade less than it was before. But in case (b) the first sample is effectively a sample from the whole  $10^{10}$ , and its composition therefore implies a high probability that the 2:1 ratio holds approximately in this, and therefore in the next 10,000, which are another sample from it. Thus in case (b) the composition of the first sample gives a considerable increase in the probability that the second will show a ratio near 2:1; in case (a) it slightly diminishes it.

Case (b) is more like what we actually meet; (a) is highly artificial.

But the fact that the inference from the first sample about the particular class sampled would be the same in both cases has been found surprising by some writers, and it seems worth while to point out that the inferences drawn about another class or a sample from one would be very different. In both cases the notion of reasonable degree of belief is involved through the notion of randomness.

It is often said that some frequency definition is implicit in the work of Bernoulli, and even of Bayes and Laplace. This seems out of the question. Bayes constructed the elaborate argument in terms of expectation of benefit to derive the product rule, which he could have written down in one line by elementary algebra if he was using the De Moivre definition. The limit definition was not stated till eighty years later, by Leslie Ellis† and Cournot,‡ and there is no mention of a limit in this part of Bayes's paper. Did Bayes go to this trouble to prove what was already obvious? Again, what can be the point of Laplace's 'equally possible' on any frequency definition? He does not mention a limit, which first appeared in the literature after his writings also. Surely Laplace's statement is meant to specify what cases he proposed to discuss; 'equally possible' is not meant to be true of all possible cases, otherwise why mention it? And if it is not always true the De Moivre definition is rejected. In his application to sampling Laplace does take the possible numbers in the population as equally possible; but this does not say that he was supposing a world population of classes with the proportions known to be uniformly distributed. I suggest indeed that the author of the *Mécanique Céleste* was much too great a man to have thought anything so ridiculous. His own statement, in the Introduction, is 'La théorie des probabilités n'est que le bon sens réduit au calcul'. His problem was simply, using the sample, to find out from it what he could about a population of otherwise unknown composition; and he said that the composition was otherwise unknown by taking the alternatives equally possible, or, as we should now say, equally probable. Similarly, Bayes gave an explicit warning again and again that the uniform assessment is to be used only when there is no information whatever about the composition of the population sampled. With such care about this point it seems remarkable that he should have omitted to say that the population was drawn from a super-population of known composition if he meant it. Such a hypothesis must be rejected on the internal evidence in Bayes's paper by any significance

† *Camb. Phil. Trans.* 8, 1843, 1-6.

‡ *Exposition de la théorie des chances et des probabilités*, Paris, 1843.

test. Similarly, it has been supposed that a limit definition is implicit in Bernoulli's theorem. But, even if the value of the limit was taken for granted, the ratio in a finite sample, however large, could mathematically still be anything from 0 to 1; the theorem would be mathematically meaningless. The ratio in a finite sample, again, has been taken as the definition of the probability, and it has been suggested that Bernoulli himself intended this to be done. Then did he construct a long and difficult mathematical argument,† showing that this ratio would be *near* the probability in the conditions considered if he was going to take it as a definition at the end? And why did he call his book *Ars Conjectandi*? I maintain that the work of the pioneers shows quite clearly that they were concerned with the construction of a consistent theory of reasonable degrees of belief, and in the cases of Bayes and Laplace with the foundations of common sense or inductive inference.

In a fairly extensive search I have not succeeded in tracing the origin of the belief that the prior probability is supposed to be derived from a known frequency. So far as I have found, Karl Pearson is the only person to have both believed anything like it and advocated the use of inverse probability. In several places he appeals to previous instances to justify the uniform assessment, which is consistent with the prior probability being, not a known frequency, but a degree of confidence based inductively on a previously observed frequency. This is entirely valid in terms of the present theory, and does not require a frequency definition. But also he sometimes says that without such previous instances the uniform assessment cannot be used, nor can any other. This, however, would make it impossible for the theory ever to find its first application. In this respect Pearson's statement is unsatisfactory, though I do not believe that even in its actual form it identifies an inferred frequency with a known one. It is, however, very difficult to understand Pearson on the point, because the development of the nature of scientific inquiry in the *Grammar of Science* often appears to be inconsistent with his statements in statistical papers, and in spite of his great achievements in introducing clarity in the *Grammar* he himself does not appear to have been influenced by them so much as might have been expected. With the doubtful exception of Pearson, however, the identification of the prior probability with a known frequency, or the statement that it must rest on one, is, so far as I have been able to

† He did not use Stirling's theorem, and his argument is much more difficult than would now be used.

trace, to be found only in the writings of opponents. I hope that this clears me from the heinous charge of originality.

8.1. The few critics of my treatment that have not proceeded by attributing to me views that I have explicitly rejected usually say that the prior probability is 'subjective' or 'mystical' and therefore meaningless,<sup>†</sup> or refer to the vagueness of previous knowledge as an indication that the prior probability cannot be uniquely assessed. On the former point, I should query whether any meaning can be attached to 'objective' without a previous analysis of the process of *finding out* what is objective. If it is done from experience it must begin with sensations, which are peculiar to the individual, and must give an account of how it is possible to proceed from the scattered sensations of an individual, including the reports of their sensations made to him by other individuals, to some set of statements that can form a possible basis of agreement for many. We must and do begin with the individual, and we never get rid of him, because every new 'objective' statement must be made by some individual and appreciated by other individuals. On the other hand, if we do not find out by experience what is objective we can do it only by imagination. One hesitates to say that critics believe that nothing but imagination is objective.

What the present theory does is to resolve the problem by making a sharp distinction between general principles, which are as impersonal as those of deductive logic, and are deliberately designed to say by themselves nothing whatever about what experience is possible, and, on the other hand, propositions that do concern experience and are in the first place always merely considered among possible alternatives. The latter are possible scientific laws; the former give rules for deciding between them by means of experience and for drawing further inferences from them. The empirical proposition is always in the first place the result of imagination. It becomes a law or an objective statement when the general rules have compared it with experience and attached a high probability to it as a result of that comparison. That is the only

<sup>†</sup> The meaning of 'metaphysics' and 'mysticism' seems to change with time. Compare the following, from J. L. Lagrange, 1760. I am indebted to Dr. F. Smithies for the reference:

'For the rest, I do not deny that it is possible, by the consideration of limiting processes from a particular point of view, to prove rigorously the principles of the differential calculus, but the kind of metaphysics which it is necessary to use in doing so is, if not contrary, at least foreign to the spirit of analysis.'

'In methods which use the infinitely little, the calculation corrects the false hypotheses automatically. . . . The error is destroyed by a second error. . . . On the other hand, Newton's method is completely rigorous.'

scientifically useful meaning of 'objectivity'. If statements about possible results of experience were included in the general principles they would lead to illegitimate *a priori* assertions about experience, and these might easily be wrong and could be disposed of, as for the first frequency definition, only by introducing contradictions.

It is argued that because  $P(p|q)$  depends on both  $p$  and  $q$  it cannot be an objective statement, since different persons with different knowledge would assess different probabilities of  $p$ . This is a confusion.  $p$  has no probability whatever of itself, any more than  $x+y$  has any particular value for given  $x$  if we do not know  $y$ . The probability of a proposition irrespective of the data has no meaning and is simply an unattainable ideal. On the other hand, two people both following the rules would arrive at the same value of  $P(p|q)$ . It is a fact that the probabilities of a proposition with respect to different data will in general differ, and people with different data will make different assessments. But this is no contradiction, but merely the recognition of an obvious fact. They will arrive at consistent assessments if they tell each other their data and follow the rules. We can know no absolute best—that would require us to have all possible knowledge. But we can give a unique and practically applicable meaning to 'the best so far as we can tell on our existing data', and that is what the theory does.

One difficulty that has possibly led to more trouble than has received explicit mention is the treatment of vague and half-forgotten empirical information. This seems to be understood in such expressions as 'uncertainty of the previous knowledge'. We have several times been led to discuss such information, and the result has always been the same: information inadequately recorded can be treated only as a suggestion of possible alternatives, and the prior probability used to express previous ignorance should still be used. The fault is not in the theory but in an imperfection of the human mind that the theory makes it possible to correct. The difference between the results of different assessments of the prior probability in the same problem is much less than the differences between those found by different statisticians that agree about little except that the prior probability must be rejected.

A prior probability used to express ignorance is *merely* the formal statement of that ignorance. It says 'I do not know' and leaves the posterior probability, if the observations are of any use for the purpose, to say 'You know now'. The statements 'I do not know  $x$ ' and 'I do not know the probability of  $x$ ' still continue to be confused. The latter is 'I do not know whether I have any information about  $x$  or not',

which differs from the former as much as  $x^4$  differs from  $x^2$ , one having been derived from  $x$  by one operation of squaring and the other by two. I should gravely doubt whether anybody approaching a set of data in the latter state of mind could possibly do anything useful with them. To speak of 'an unknown prior probability' involves either this confusion or the identification of the prior probability with a world-frequency, and no coherent theory can be made until we are rid of both.

The confusion may arise partly from the fact that probability statements are sentences in the indicative mood. Thus the question 'Is Mr. Smith at home?' can be expressed by three sentences in the indicative mood:

I do not know whether Mr. Smith is at home.

I want to know whether Mr. Smith is at home.

I believe that you know whether Mr. Smith is at home.

These three sentences contain the whole content of the question, and the difference from 'Mr. Smith is at home' is expressed by a transposition of subject and verb and, in print, a symbol called a question-mark. The situation implied in these three statements is so common that a special symbolism has been introduced into language to express it. The prior probability statement is the first. The second is, in a scientific problem, indicated sufficiently by our willingness to undertake the work of finding the answer; it is a statement of a wish and is not a probability statement. The third is a probability statement of higher order; and all this is done in speech by a transposition. Yet people continue to question whether degrees of knowledge can be expressed in symbols. What the prior probability does, in fact, is to state clearly what question is being asked, more clearly than ordinary language is capable of doing. And I suggest that this is no mean achievement. Many will support me when I say that 90 per cent. of the thought in a scientific investigation goes in the preliminary framing of the question; once it is clearly stated, the method of answering it is usually obvious, laborious perhaps, but straightforward. Consider, for instance, the work of G. I. Taylor and H. Quinney on the plasticity of copper,<sup>†</sup> to decide whether the difference between the largest and smallest principal stresses at a point, or the Mises function, which is a symmetrical function of the three principal stresses, afforded the correct criterion for the start of flow. It was known that different specimens of the material differed more than the difference between the criteria

<sup>†</sup> *Phil. Trans. A*, 230, 1932, 323-62.

would be. Hence to answer the question it was necessary to eliminate this variation by working on the same specimen throughout. But then something that would differ according to the criterion had still to be found. They showed that if tension  $P$  and shear stress  $Q$  were applied simultaneously, the former directly, the latter by torsion, the Mises criterion would give flow at a constant value of  $P^2 + 3Q^2$ , the stress-difference at a constant value of  $P^2 + 4Q^2$ . Here at last was an answerable question clearly stated. The suggested experiment needed care and skill, but not much more; the brilliance was in asking the right question. It would be easy to give a long list of papers that cannot answer the question that they claim to answer, simply because insufficient attention has been given to whether the data are suited to decide between the possible alternatives.

Part of the objection to probability as a primitive notion is connected with the belief that everything is vague until it is defined in words. Such a belief omits to recognize that some things are perfectly intelligible before any definition is available. To try to define such things can result only in defining them in terms of something less immediately intelligible and failing to give account of established laws. For instance, observed colours are found to be associated with different measured wave-lengths. This led to the idea that colour should be *defined* in terms of the wave-length and the sensory impression rejected. This was vigorously advocated; but had it been acted upon nobody would have been able to say that a thing was red until he had actually set up a spectroscope and measured the wave-length of the radiation coming from it. Not even the persons with the facilities for doing it would act on the principle. What the recommendation does is to reject an important means of investigation, and the empirical relation between colour and wave-length. The behaviourist psychologists reject consciousness and thought except so far as they can define them in terms of certain minute movements in the throat that go on when the person says he is thinking. Consequently, in their system, there are two alternatives. (1) A man has no way of knowing whether or what he is thinking except by observing these movements. Many people manage very well without it. (2) He may admit his own consciousness but reject other people's. That is solipsism, and no two solipsists can understand each other and agree. Eddington, finding the fundamental laws of physics symmetrical with regard to past and future, searches for something that does vary in one direction with time and finds entropy; and therefore defines the order of increasing time as that of



increasing entropy. Consequently he could not know that he wrote the *Relativity Theory of Protons and Electrons* after he discovered the mass-luminosity relation except by measuring the entropy of the universe on the two occasions. It all seems very difficult. Bertrand Russell, who cannot be accused of shirking the logical consequences of his postulates, or of refusing to change the postulates when the consequences are intolerable, has arrived at the conclusion:† ‘*Things are those series of aspects which obey the laws of physics.*’ That such series exist is an empirical fact, which constitutes the verifiability of physics.’ Much of what passes for modern theoretical physics consists in the application of the first sentence while forgetting the second. To be a practical definition it must refer to the laws already known, not to the aggregate of all laws. In the former sense it is a possible rule for progress; in the latter it is a mere counsel of perfection. But in the former sense the fact that series have been found to fit the laws is equivalent to saying that laws have been found to fit the aspects. Russell, be it noted, does not define an aspect, but merely gives a rule about what aspects are to be grouped in a series to constitute a thing; and the second sentence recognizes that a possible law must be rejected if no series of aspects can be found that conform to it.

Definitions add clarity when something new is defined in terms of something already understood; but to define anything already recognizable is merely to throw valuable information into the wastepaper basket. All that can be done is to point to instances where the phenomenon in question arises, in order to enable the reader to recognize what is being talked about by comparison with his own mental processes and sensations.

W. E. Johnson‡ puts the point even more strongly. He remarks that some things are ‘so generally and universally understood that it would be mere intellectual dishonesty to ask for a definition’.

**8.2.** We can never, formally, rule out the possibility that some new explanation may be suggested of any set of experimental facts. But we have seen that in many cases this does not matter, by 1.6. Once a law has attained a high probability it can be used for inference irrespective of its explanation. If an explanation also accounts for several other laws, so much the better; there is more for any alternative to explain before it can be said to be as satisfactory as the existing one. The question of an alternative becomes effective only when (1) it accounts

† *Our Knowledge of the External World*, 1914, p. 110.

‡ *Logic*, 1, 106.

for most or all of the evidence explained by the first, (2) it suggests a specific phenomenon that would differ according to which is right. The decision can then be made in accordance with our principles. This is the answer returned by the theory of probability to the logical difficulty of the Undistributed Middle, or the neglect of an unforeseen alternative. The use for inference is valid so long as it involves only the use of laws that have already been established inductively, because the laws are in a stronger position than any explanation could possibly be. When an explanation is used and applied to predict laws, these require test; but now the possible alternative explanations are severely limited by the fact that they must agree with the laws already known. Incidentally, this meets a possible difficulty with the rule that all suggestions have the same prior probability, no matter who makes them. The layman in a subject may be admitted as capable of making a good guess, but it is extremely hard for him to make a guess that is not contradicted by evidence already known.

This also answers the problem of 'scientific caution'. Everybody agrees on the need for caution, but different people, or even the same person on different occasions, may have entirely different opinions on what caution means. I suggest that the answer is that results should always be presented so that they will be of the maximum use in future work. That involves, for pure estimation, a statement of a location parameter and its standard error. But it can never be guaranteed that no modification in a law will ever need to be considered; and a possible systematic error of observation needs positive evidence for its existence just as any other modification does. To assert in advance any kind of departure from the suggested law is a reckless statement, irrespective of whether the departure considered is a systematic error of observation or a 'physical' effect that the physicist considers more interesting. In both cases the information should be presented so that a significance test can be applied when suitable evidence is available; and this implies giving the estimated value, the standard error, and the number of observations. There is no excuse whatever for omitting to give a properly determined standard error. It is a necessity in stating the accuracy of any interpretation of the data, if the law is right; if the law is wrong, it is necessary to the discovery that it is wrong. All statisticians will agree with me here, but my own applications are mostly in subjects where the need is still very inadequately appreciated. Again, the best way of finding out whether a law is wrong is to apply it as far as possible beyond the original data, and the same applies to

any suggested explanation. But if we have not a determination of the standard errors of the parameters in the law we have no way of saying whether any discrepancy found is genuine or could be removed by a permissible readjustment of the parameters, with a corresponding improvement in their accuracy. The usual reason given for the omission is that there may be some other source of error and that the statement of a standard error expresses a claim of an accuracy that future events may not justify. This rests on a complete failure to understand the nature of induction. It is essential to the possibility of induction that we shall be prepared for occasional wrong decisions; to require finality is to deny the possibility of scientific inquiry at all. The argument, however, does not prevent its users from asserting systematic differences when the estimates agree within the amounts indicated by the standard errors, supposing these genuine, or from denying them when they are flagrant. What we should do is (1) always to draw the most probable inference from the data available, (2) to recognize that with the best intentions on our part the most probable inference may turn out to be wrong when other data become available, (3) to present our information in such a form that, if we do make mistakes, they can be found out. This can be done by a consistent process, and should not be confused with guesswork about other possible effects before there is any evidence for their existence or any estimate of their amount.

8.3. The situation with regard to alternative explanations mentioned above actually existed for a long time in relation to the quantum theory. The quantum explanation seemed to be demanded by the distribution of black-body radiation and by the photo-electric effect; it seemed to be denied by the phenomena of interference, notably by G. I. Taylor's experiment,<sup>†</sup> which obtained interference patterns under illumination of intensity so low that it was highly improbable that there would ever be two quanta inside the apparatus at once. The quantum theory and the continuous emission theory both accounted for one set of facts, but each, in its existing form, was inconsistent with the facts explained by the other. The proper conclusion was that both explanations were wrong, and that either some new explanation must be sought or the sets of data recognized as unrelated. But meanwhile, physicists based their predictions on the laws; in types of phenomena that had been found predictable by quantum methods, they made their predictions by quantum methods; in phenomena of interference they

<sup>†</sup> *Proc. Camb. Phil. Soc.* **15**, 1909, 114-15.

made predictions by assuming continuous wave trains. Thus what they really did was to proceed by induction from the laws established empirically. This was a valid process and did not require the assertion of any particular explanation of the laws, the latter being entirely subsidiary.

The present position of the quantum theory illustrates another point in relation to the theory of probability. There are three main quantum theories; but all make the same predictions and for, it may be, the first time in the history of physics, the exponents are willing to accept the situation and even on occasion to use one another's methods. The theories themselves are not the same, and indeed each contains reference to things that have no meaning on another. The treatment of them as equivalent refers only to the observable results predicted, and not to their actual content. It recognizes that as long as theories lead to the same predictions they are not different theories, but merely different ways of saying the same thing. The differences are relegated to metaphysics. But this is a complete abandonment of naïve realism, in which the things with 'physical reality' would be those contained in the explanations, and no others. It does not matter, for instance, whether an electron is a point charge with an exact position that we do not quite know, or a volume distribution rather fuzzy at the edges, or whether the position of the electron is intrinsically meaningless in the sense that it cannot be expressed in terms of three Cartesian coordinates at all. This attitude is precisely what is reached here; the essential thing is the representation of the probability distribution of observable events, and therefore the forms of laws and the values of parameters in them. Questions that cannot be decided by means of observation are best left alone until some way of answering them suggests itself.

8.4. The modern quantum theories, like the relativity theories, suffer from a confusion in the use of the term 'the rejection of unobservables'. 'Unobservable' is a legacy from naïve realism. An observation, strictly, is only a sensation. Nobody means that we should reject everything but sensations. But as soon as we go beyond sensations we are making inferences. When we say that we have observed an object we mean that we have had a series of sensations that are coordinated by imagining or postulating an object with assigned properties, and that to continue to do so will probably lead us to a correct prediction of other groups of sensations. "To observe an object" is merely an idiomatic shorthand way of writing this; what we really observe is a series of

patches of colour of various shapes, and whether these are correctly located in our minds or where we suppose the object to be must be left to philosophers. But in naïve realism it is taken for granted that we do observe the object and that the patches of colour are 'subjective' and not respectable; and this puts the cart before the horse because except through the latter there is no way of finding out anything about the object at all. The acceptance of an object with its properties depends on the verification of the inferences that it leads to; that is, it is required that our sensations without it, or if it had different properties, would be different from what they have actually been. Hence the verifiable content can be stated entirely in terms of parameters in laws connecting sensations. This is dealt with completely by the theory of probability, and for purposes of inference the laws are all we want. If we restrict ourselves to the inference of future sensations the concept has done its work and serves no other purpose. This would be a possible idealist attitude. If we are realists and think that our concepts have counterparts in an external world (subject to the critical realist's willingness to change his mind if necessary), we may consider the law as a justification of the reality of the concept. But observability of a concept can mean nothing but the statement that it suggests new parameters in laws connecting sensations, and that the need for these parameters is supported by a significance test. Thus the theory of probability takes the rejection of observables in its stride. It gives an answer to the question whether any parameter is more probably present than not, given the actual data. To consider further data that we have not is sheer waste of time. We do not say that so-and-so *must be unobservable*; we say that, with the information at our disposal, it *is unobserved*, and that if we try to take it into account we shall probably lose accuracy. To say that it must be unobservable would be illegitimate; it would be either an *a priori* statement leading to inferences about observations or an induction claiming deductive certainty.†

The principle really seems to have arisen from a confusion of three possible statements of the 'economy of hypotheses'. (1) In developing a logic, as in *Principia Mathematica*, the number of postulates is reduced to a minimum, though some results that appear as theorems appear equally obvious intuitively. The reasons for this procedure have been discussed under rule 6 of Chapter I. (2) Parameters in a law that make

† Cf. H. Dingle, *Nature*, 141, 1938, 21-8. This is an admirable statement of the logical position of the principle, except for the omission to consider any realism but naïve realism.

no contribution to the results of any observation can be eliminated mathematically, leaving the observations to be described only in terms of the relevant parameters. When this is done an economy of statement may be achieved (possibly at the cost of increased complexity of mathematical form), but there is no improvement in representing either present or future observations, since either form will say precisely the same thing about both. (3) The third is the simplicity postulate as used in the present theory, which leads to the restatement of Ockham's principle in the form 'Variation must be taken as random until there is positive evidence to the contrary'. This is the principle that we actually need. The second principle is always a pure tautology; but in the usual statement it becomes the 'rejection of unobservables' and is used to deny the relevance of any variable not yet considered. It then becomes an *a priori* statement that future observations *must* follow certain laws, whatever the observations may say. Such an inference into the future must be an inductive inference based on probability, because it is logically possible that the observations may disagree with prediction. The third principle deals with such inferences, but the attempt to use the second involves a logical fallacy.

Now I maintain that whatever has been said on the matter, the rejection of unobservables in the form stated has never led to a single constructive advance, and that in spite of the reluctance of modern physicists to pay any serious attention to the problem of induction, what they have done is to use induction and then confuse it with deduction. Relativity, up to 1920 or so at any rate, did not involve any new parameters; the velocity of light, the constant of gravity, the mass of the sun, and so on, were all required by previous theories. It made changes in the laws but left them expressed in terms of the same parameters. The reason for abandoning the old theory was not that it involved unobservables such as absolute velocity or simultaneity; it was that this theory made positive predictions, such as the one sought for in the Michelson-Morley experiment, which turned out to be in disagreement with observation. The rejection of absolute velocity was not *a priori*; what was done in the special theory of relativity was to alter the laws of measurement and light so that they would agree with observation. The general theory, in its original form, was obtained by a natural analogy with Newtonian dynamics. The coefficients  $g_{ij}$ , in what seemed to be the natural extension of the special theory to take gravitational effects into account, were seen to play the part of the Newtonian potential  $U$ . Far from matter all second derivatives of

the latter vanish; near to matter the contracted Cartesian tensor  $\nabla^2 U$  vanishes, but the separate components do not; inside matter  $\nabla^2 U$  does not vanish, but has a simple relation to the density. Einstein proceeded by analogy. He found a second-order tensor that should vanish far from matter, contracted it to get the differential equations satisfied near matter, and said that these equations will be modified inside matter. Given, what was already established, that the Euclid-Newton system needed modification, this was the natural procedure to try. But it is a suggestion, not an *a priori* necessity. On this point one may refer to Eddington, writing just before the 1919 eclipse expeditions:† 'The present eclipse expeditions may for the first time demonstrate the weight of light; or they may confirm Einstein's weird theory of non-Euclidean space; or they may lead to a result of yet more far-reaching consequences—no deflexion.' The first alternative refers to the Newtonian deflexion, which would be half Einstein's. That was Eddington's position before the observational result; Einstein's theory stood to him as the theory of probability says that it should, as a serious possibility needing test, not as demonstrable by general principles without reference to observation. In other words, Eddington at the proper time agreed with me; his later emphasis on the mathematical necessity of Einstein's theory is a case of 'forgetting the base degrees'. The correctness of Einstein's law rests on the fact that it requires no new parameters and gives agreement with observation where the alternatives fail. Insistence on the alleged philosophical grounds for it has led to their being challenged, and to a tragic neglect of the observational basis. The latter is, in fact, appreciably stronger than is provided by the mere verification, as I showed in chapters vii–ix of *Scientific Inference*. Starting entirely from observed data and proceeding by generalization of laws, introducing new parameters only when observation showed them to be necessary, I showed that it was possible by successive approximation to build up Euclidean mensuration, Newtonian dynamics, and the special and general theories of relativity; and that the form of Einstein's  $ds^2$  is completely determined near the sun by observation alone. No further hypothesis is needed, and some of those made by Einstein are replaced by others more closely related to laws already adopted or by experimental facts. The linearity of the transformation of coordinates in the special theory, for instance, need not be assumed. It can be proved from the constant measured velocity of light and the natural extension of Newton's first law, that an unaccelerated particle

† *The Observatory*, March 1919, p. 122.

in one inertial frame must be unaccelerated in another. The object of the work was to see whether the observed agreement could be regarded as accidental, that is, whether any other possible laws (Newton's in particular) could have given the same results in the range of magnitude available; and it was found that no other form would explain on Newton's theory a fact not explained on Einstein's without leading to contradictions elsewhere. For instance, the excess motion of the perihelion of Mercury had been known for ages to be explicable by the attraction of an oblate distribution of matter around the sun, such as was seen in the zodiacal light; and with a suitable inclination of the axis such matter could also explain the excess motion of the node of Venus, which is not explicable on Einstein's theory and is too large to be regarded as random error. To explain it by gravitation would require enough matter to upset the agreement for the perihelion of Mercury. Similarly, it was suggested, I believe by Professor H. F. Newall, that the eclipse deflexion could be explained by the refraction of matter near the sun. But such Newtonian explanations led to estimates of the amount of matter needed, and according as it was solid or gaseous the amount of light it would scatter could be estimated. It was found that the visible scattered light did not correspond to more than an insignificant fraction of what would be implied by the Newtonian explanation.† Using some more recent data I find a larger discrepancy. Hence there is no Newtonian explanation in sight for either the perihelion of Mercury, the node of Venus, or the eclipse displacement; while Einstein's law explains the first and third. The node of Venus is not evidence for Newton's law, because this does not explain it either. This discrepancy is apparently significant, but what it signifies is not clear; it may represent some systematic error of observation or internal correlation of the errors, though these have not been adequately tested. What is quite clear, however, is that it is irrelevant to the decision between the two laws of gravitation. So far as any law can be proved by observation (and no law can be proved at all in any other way), Einstein's law is proved within the solar system.

The rejection of unobservables in the quantum theory seems to be a mere spring-cleaning and to be correctly placed under the second of the above principles. The older theories involved many unobservable quantities, and left many observable ones uncoordinated. It had become impossible to see the wood for the trees on account of the complications of the concepts, and the postulates led to results inconsistent

† *M.N.R.A.S.* 80, 1919, 138-54.



with observation. The modern quantum theories have begun by direct and successful attempts to coordinate what we know, without attending to the details of any deeper interpretation, and this was right as a matter of mathematical convenience. But it is no more a rule for positive discovery than the fact that a gardener weeds his plot before sowing his seed. The important forward step did not come from the rejection of unobservables but from the subsequent recognition of formal relations. These relations are not inferred from a principle that so-and-so must be unobservable—and indeed they are full of new unobservables of their own, which have to be eliminated before anything verifiable is reached. They are guessed by analogy with Newtonian dynamics and asserted because their consequences agree with observation, just like Einstein's law of gravitation.

The most elaborate use of the form of the rejection of observables criticized on p. 385 is to be found in the works of Eddington, culminating in his statement that all the fundamental laws and constants of physics can be predicted from purely epistemological considerations. Some comments on his conclusion are given in 5.64; a criticism of his general point of view in the *Philosophical Magazine* paper cited there.

A warning is needed that the frequent use of the word 'probability' in works on quantum theory is no guarantee that the numbers referred to are probabilities in any sense or satisfy the laws of probability, and that there is reason to suppose that the probability interpretation of wave mechanics leads to the conclusion that quantum theory is deterministic in exactly the same sense as classical mechanics.†

8.5. Criticism of fallacious logic is usually treated as captious, on the grounds that the methods criticized have delivered the goods. It is not considered a matter of importance to physics whether the arguments are right so long as they somehow give the right answer at the end. But the methods have not delivered the goods. The chief advances in modern physics were not achieved by the rejection of unobservables or by any other alleged general mathematical principle. They were achieved by the method of Euclid and Newton: to state a set of hypotheses, work out their consequences, and assert them if they accounted for most of the outstanding variation. The method was inductive, and the claim that the results were obtained in any other way is contrary to history. The insistence on the mathematical argument as a proof, in turn, invites challenge on grounds of logic; either it is

† Cf. *Phil. Mag.* (7), 33, 1942, 815–31.

important or it is not. If it is, it must be prepared to meet logical criticism by a logical answer; if it is not, it should be dropped and cease to make bad logic an essential part of what is supposed to be mathematics. Above all, it should cease to obstruct the development of an adequate theory of induction.

Reasoning and observation are two different faculties, and it is important to keep them separate, as far as possible, and to separate them as well as we can if the information presented to us is in such a form that they have already been mixed. If this is not done we may find ourselves in the position of saying that the argument is right and therefore we do not need observations to test whether we have overlooked anything; or that the argument leads to results agreeing with observation and therefore must be right however many mistakes are found within it. Many modern examples of both could be found. The following one, though not exactly recent, is an interesting illustration of how attention to the details of an argument has actually led to constructive results. Laplace in his calculation of perturbations had shown that the eccentricity of the earth's orbit should be systematically diminishing. This affects the disturbance of the moon by the sun, and leads to the result that the moon's distance should be decreasing, and its rate of revolution about the earth increasing. This would alter the calculated times of ancient eclipses, and recorded observations of them showed that such an effect was required. Laplace gave only the first term of the series representing it, but this was near enough to the observed value for Plana, Damoiseau, and Hansen to develop the matter and include further terms. The agreement at this point seemed entirely satisfactory. J. C. Adams, however, worked out the theory afresh† and found that several neglected terms mounted up. The first two coefficients of the series in powers of  $m$ , where  $m$  is the ratio of the mean motions, are  $\frac{3}{2}m^2 - \frac{3771}{64}m^4$ , whereas Plana had got  $-\frac{2187}{128}m^4$  for the second. On account of this enormous numerical coefficient the calculated value of the secular acceleration was practically halved, and the agreement with observation was destroyed. Adams's result was confirmed by Delaunay and several other dynamical astronomers, who obtained further terms. But Pontécoulant said that if the result of Adams were admitted it would 'call in question what was regarded as settled, and would throw doubt on the merit of one of the most beautiful discoveries of the illustrious author of the *Mécanique céleste*'. Le Verrier wrote: 'Pour un astronome, la première condition est que ses théories satis-

† *Phil. Trans.* 143, 1853, 397-406; see also several of his collected papers.

fassent les observations. Or la théorie de M. Hansen les représente toutes, et l'on prouve à M. Delaunay qu'avec ses formules on ne saurait y parvenir. Nous conservons donc des doutes et plus que des doutes sur les formules de M. Delaunay. Très certainement la vérité est du côté de M. Hansen.' Thus the mathematics of Adams and Delaunay was to be judged, not by whether the results followed from the equations of dynamics, but by whether they agreed with observation; if the results disagreed with observation there must be a mistake in the mathematics. J. W. L. Glaisher remarks in his biographical notice:† 'It is curious that it should have been possible for so much difference of opinion to exist upon a matter relating only to pure mathematics, and with which all the combatants were fully qualified to deal, as is clearly shown by their previous publications.' What happened, in fact, was that Adams's result was so thoroughly confirmed by different methods and different investigators that it had to be accepted and the discrepancy admitted. But the result was not purely destructive. What it did was to direct attention to the matter afresh and to lead to the theory of tidal friction in a long series of papers by Sir G. H. Darwin;‡ and at the present time this appears to give quite satisfactory quantitative agreement with observation,§ and a large number of constructive results about the remote past and future of the solar system, which could never have been considered at all if Plana's result had stood unquestioned.

The use of the word 'theory' in several different senses is perhaps responsible for a good deal of confusion. What I prefer to call an 'explanation' consists of several parts: first, a statement of hypotheses; secondly, the systematic development of their consequences; thirdly, the comparison of those consequences with observation. It still sometimes happens, as in some passages just quoted, that the fact that the alleged consequences agree with some observations is a proof both that the hypotheses are right and that the intermediate steps have been correctly worked out. What is liable to be true is that the intermediate development involves numerous begged questions, the answers having been chosen so as to agree with observation and not because they are consequences of what has gone before; and that the correct working out of the consequences leads to results disagreeing with the very observations that the theory is said to explain. In such cases the hypotheses are disproved. Further, it is open to anybody to work out other conse-

† Adams, *Collected Works*, p. xxxviii.

‡ *Scientific Papers*, vol. 2.

§ G. I. Taylor, *Phil. Trans. A*, 220, 1919, 1-33; Jeffreys, *ibid.* 221, 1920, 239-64; *The Earth*, 1929, ch. xiv.

quences of the hypotheses and to see whether these agree with observation, and if they do not, to suggest a different set of hypotheses. That is how science advances. There are some current 'theories' that, when divested of begged questions, reduce to the non-controversial statement, 'Here are some facts and there may be some relation between them'.

8.6. To recapitulate the main postulates of the present system, we have first the main principle that the ordinary common-sense notion of probability is capable of consistent treatment. Other theories can deny the consistency, but cannot help using the notion. We have also Axiom 4, which implies that there is no inconsistency in using the addition rule. The rule as it stands is a convention, since other rules consistent with the axioms would be possible and would lead to putting probabilities in the same order, and all could be compared with a standard obtained by considering balls in a bag. Thus the numerical assessment merely specifies the rules of a language capable of going into more detail than ordinary language. A generalization of the product rule may be needed, justified by the principle adopted in *Principia Mathematica* that in constructing a logic the postulates should be taken in their most general form. These postulates are required in all theories. The principle of inverse probability is a theorem. The prior probabilities needed to express initial ignorance of the value of a quantity to be estimated, where there is nothing to call special attention to a particular value, are given by an invariance theory that leads to equivalent results for transformations of the parameters, combined with some rules of irrelevance to the effect that the actual values of certain parameters, especially scale parameters, tell us nothing about those of certain others. Where a question of significance arises, that is, where previous considerations call attention to some particular value, half the prior probability is concentrated at that value. This is the simplicity postulate. It needs some elaboration when several parameters arise for consideration simultaneously.

The main results are: (1) a proof independent of limiting processes that the whole information contained in the observations with respect to the hypotheses under test is contained in the likelihood, and that where sufficient statistics exist other functions of the observations are irrelevant; (2) a development of pure estimation processes without further hypothesis; (3) a general theory of significance tests, which allows any hypothesis to be tested provided only that it is sufficiently clearly stated to be of any use if it is true, declares no empirical hypothesis

to be certain or false *a priori*, does not require the introduction of the *P* integral to avoid results in contradiction with common sense, and leads to a solution of the estimation problem as a by-product of the significance test instead of as a separate problem based on contradictory hypotheses; (4) arising out of this, an account of how in certain conditions a law can reach a high probability and inferences from it be treated as deductive in an approximate treatment. It thus makes it possible to test laws by observation, without making either the unnecessary assumption that laws can be found to fit the observations exactly, or the false one that laws known to us at present do; thus it gives a formal account of the actual process of learning. Further, it solves the problem of the rejection of unobservables, replacing a useless mathematical platitude by a practical criterion; removes the paradoxical appearance of the uncertainty principle; meets the logical difficulty of the undistributed middle; and gives intelligible meanings to 'scientific caution' and the notion of 'objectivity'.

Comment was made in Chapter I on the fact that a formal and consistent theory of inductive processes cannot represent the operation of every human mind in detail; it will represent an ideal mind, but it will also help the actual mind to approximate to that ideal. We have had occasion sometimes to call attention to special imperfections, notably: (1) wish-fulfilment, expressed sometimes in an exaggerated lenience towards one's own hypotheses, sometimes in a belief that things can be proved in terms of ordinary mathematics and deductive logic when in their very nature they cannot be, and an appearance of such a proof is simply a proof that there must be a mistake in it; (2) imperfect memory, which can be treated merely as a suggestion of alternatives but not as a contribution of observational information when the matter is brought up for formal consideration; (3) failure to think of the right empirical hypothesis at the time when data are first available to test it; (4) limitations of time or industriousness that make us content with approximations. The existence of these is no argument against the theory; but the theory will provide a standard of comparison for them in psychological studies; psychology is admitted as a valid science to the same standards as any other.

The human mind has also a tendency to exaggerate the differences between familiar things and overlook the resemblances. Let us recall the reply of Dr. Jervis to a lady who had asked whether Dr. Thorndyke was 'at all human'.† "He is entirely human," I replied, "the accepted

† R. Austin Freeman, *John Thorndyke's Cases*, p. 60.

test of humanity being, as I understand, the habitual adoption of the erect posture in locomotion, and the relative position of the end of the thumb——”

“I don’t mean that,” interrupted Mrs. Haldean. “I mean human in things that matter.”

“I think those things matter,” I rejoined. “Consider, Mrs. Haldean, what would happen if my learned colleague were to be seen in wig and gown, walking towards the Law Courts in any posture other than the erect. It would be a public scandal.”

We have, of course, the words ‘person’ and ‘human’, which can apply to any member of the species. But though we have six or seven words to describe different sexes and ages of the species *Canis familiaris*, *Bos taurus*, *Equus caballus*, we have no standard word that can apply to any individual of either.† The real reason for the difficulty in the understanding of the theory of probability is, I think, that the fundamental ideas and general principles are so familiar that ordinary language has overlooked them, and when they are stated it is immediately taken for granted that they *must* mean something too complicated for ordinary language, and a search is made for something to satisfy this condition. The truth is that they are too simple for ordinary language, and the customary approach renders any understanding impossible.

8.7. We now return to the question of realism versus idealism. The question is whether the theory leads to any decision between them. Nothing in the theory depends on the acceptance of one or the other, and to arrive at a decision in terms of it we must point to some observable fact that would be more probable on one than on the other. Both are admissible hypotheses and we must take their prior probabilities as  $\frac{1}{2}$ . We see that solipsism, the extreme form of idealism, can be rejected by the theory. If other people had not minds something like my own it would be very improbable that their behaviour would resemble mine as much as it does. The belief in a material world is on a different footing, since while I seem to be immediately aware of my own personality, any object, even my own body, is known to me only through sensations. If I was an idealist I should say that I had invented it to give a convenient way of describing my sensations (past, present, and future, so far as they can be inferred, since we are not considering the rejection of induction). A realist would say that he

† Curiously, the infantile ‘bow-wow’, ‘moo-moo’, ‘gee-gee’ can apply to any member of the respective species. The loss of general words has taken place in acquiring adult language.

meant something more than that, but it is very difficult to say just what. Personally I believe that in studying seismology I am finding out something about the interior of the earth and not merely making predictions about future observations. But in either case the rival hypotheses could be tested only through the sensations predicted from them; and the properties that the idealist would assign by convention to his imaginary objects would be such as to lead to exactly the same predictions as those that the realist would postulate of the objects that he supposes real. Thus the theory of probability makes no decision whatever between critical realism and critical idealism, if the latter is taken as admitting other personalities; both have probability  $\frac{1}{2}$ , and there appears to be no type of evidence that could alter this. An attempt to support idealism has been made by saying that realism involves an extra hypothesis and should therefore be rejected if evidence for it is not available. This appeal to the economy of hypotheses is not valid, however. It only justifies the omission to assert realism; that is, it still leaves us in the position 'either idealism or realism is true' but agreeing to say no more about it. The denial of the extra hypothesis is just as much a hypothesis as its assertion. The conclusion we reach, therefore, is that there are forms both of realism and of idealism that would be scientifically tenable, that scientific method cannot decide between them, and that it doesn't matter anyhow. But neither of them is the form of realism or idealism usually advocated. Realism has the advantage that language has been created by realists, and mostly very naïve ones at that; we have enormous possibilities of describing the inferred properties of objects, but very meagre ones of describing the directly known ones of sensations; 'probability' is a word of five syllables, whereas the use of the notion dates from a time when one would be beyond our powers. So the idealist must either do his best with realist language or make a new one, and not much has been done in the latter direction.

Questions like these, that cannot be answered by scientific means, may be called metaphysical. (I do not regard this as a mere term of abuse.) Another is the distinction between religion and materialism. A materialist can hold that all biological phenomena, including evolution, are due to physical and chemical causes; he cannot state just why a Nautilus evolved into an ammonite, nor why an ammonite did not evolve back into a Nautilus, but he cannot be refuted on this ground because he can always appeal to the fact that the consequences of the laws have not yet been fully worked out and in any case there are

presumably physical laws that are not yet known. Bishop Barnes can accept evolution and reject the account of creation in Genesis, and hold that evolution is the actual way the Creator creates species and that He laid down the physical laws in the first place. To him the discovery of scientific laws is the discovery of something about how the Creator works. Equally he cannot be refuted; it would be impossible to produce any piece of observational evidence that could not be dealt with in this way. His view and the materialist's are scientifically equally tenable; the choice between them is apparently a matter of what one wishes to believe and not of evidence. In spite of G. K. Chesterton's opinion to the contrary, many people do find an emotional satisfaction in materialism. The opposition often alleged between religion and science arises only when religion ceases to be religion and becomes bad science. Actually they are mutually irrelevant. This is fortunate; it enables, for instance, both the Jesuit Seismological Association and Soviet Russia to produce good seismological observations. Similarly for the distinction between free will and determinism. The determinist can always say 'it is predestined what I shall do; so there is only one course open to me; here goes!' The *Arabian Nights* may be studied for examples.

8.8. The present theory does not justify induction. I do not consider justification necessary or possible; what the theory does is to provide rules for consistency. A prediction is never in the form 'so-and-so will happen'. At the best it is of the form 'it is reasonable to be highly confident that it will happen'. This may be disappointing, but in the last resort that is all that we can say. The former statement is a fallacious claim to deductive certainty; the latter is attainable by a consistent process. In this sense we can justify particular applications, and it is enough.



## APPENDIX

### TABLES OF $K$

WE have defined 
$$K = \frac{P(q|\theta H)}{P(q'|\theta H)},$$

where  $q$  is the null hypothesis,  $q'$  the alternative,  $H$  the previous information, and  $\theta$  the observational evidence. We take the standard case where  $q$  and  $q'$  are equally probable given  $H$ . In most of our problems we have asymptotic approximations to  $K$  when the number of observations is large. We do not need  $K$  with much accuracy. Its importance is that if  $K > 1$  the null hypothesis is supported by the evidence; if  $K$  is much less than 1 the null hypothesis may be rejected. But  $K$  is not a physical magnitude. Its function is to grade the decisiveness of the evidence. It makes little difference to the null hypothesis whether the odds are 10 to 1 or 100 to 1 against it, and in practice no difference at all whether they are  $10^4$  or  $10^{10}$  to 1 against it. In any case whatever alternative is most strongly supported will be set up as the hypothesis for use until further notice. The tables give values of  $\chi^2$ ,  $t$ , or  $z$  for  $K = 1, 10^{-1/2}, 10^{-1}, 10^{-3/2}, 10^{-2}$ . The last will be regarded as a limit for unconditional rejection of the null hypothesis.  $K = 10^{-1/2}$  represents only about 3 to 1 odds, and would be hardly worth mentioning in support of a new discovery. It is at  $K = 10^{-1}$  and less that we can have strong confidence that a result will survive future investigation. We may group the values into grades, as follows.

Grade 0.  $K > 1$ . Null hypothesis supported.

Grade 1.  $1 > K > 10^{-1/2}$ . Evidence against  $q$ , but not worth more than a bare mention.

Grade 2.  $10^{-1/2} > K > 10^{-1}$ . Evidence against  $q$  substantial.

Grade 3.  $10^{-1} > K > 10^{-3/2}$ . Evidence against  $q$  strong.

Grade 4.  $10^{-3/2} > K > 10^{-2}$ . Evidence against  $q$  very strong.

Grade 5.  $10^{-2} > K$ . Evidence against  $q$  decisive.

Any significance test must depend on at least two variables, the number of observations and the estimate of the new parameter (more usually the ratio of the latter to its estimated standard error). Consequently any table of  $K$  must be a table of at least double entry. In the tables I have taken those tests where  $K$  depends on not more than two variables. In most of each table the computations were based on the asymptotic formula, values for small numbers of observations being

separately computed from the exact formula. Accuracy of a few per cent. was considered sufficient, since it will seldom matter appreciably to further procedure if  $K$  is wrong by as much as a factor of 3.

It is clear from the tables how accurately it is worth while to do the reduction of a given set of observations. Consecutive values of  $\chi^2$  or  $t^2$  for given  $\nu$  usually differ by at least 10 per cent., often by 20 per cent. or more. If we get  $\chi^2$  or  $t^2$  right to 5 or 10 per cent. we shall in practice be near enough, and this implies that the work should be right to about 5 per cent. of the standard error. Hence as a general rule we should work to an accuracy of two figures in the standard error. More will only increase labour to no useful purpose; fewer will be liable to put estimates two grades wrong. For instance, suppose that an estimate is quoted as  $4 \pm 2$  from 200 observations, to be tested by Table III. This might mean any of the following:

	$t^2$	Grade
$4.5 \pm 2.5$	9.0	2
$3.5 \pm 2.5$	1.96	0
$4.0 \pm 2.0$	4.0	0

Similarly,  $5 \pm 2$  from 200 observations might mean any of:

	$t^2$	Grade
$4.5 \pm 2.5$	3.24	0
$5.0 \pm 2.0$	6.25	1
$4.5 \pm 1.5$	9.0	2
$5.0 \pm 1.5$	11.1	3
$5.5 \pm 1.5$	13.4	4

The practice of giving only one figure in the uncertainty must therefore be definitely condemned, but there is no apparent advantage in giving more than two. Similarly, minor correcting factors in  $K$  that do not reach 2 can be dropped, since decisions that depend on them will be highly doubtful in any case.

It will be noticed in Table I that for small numbers of observations  $K = 1$  is at not much over the standard error. This is rather surprising, but becomes less so when we consider the values of  $K$  in testing an even chance from samples of 5 and 6.

$x$	$y$	$\chi^2$	$K$	$x$	$y$	$\chi^2$	$K$
5	0	5.0	$\frac{3}{16}$	6	0	6.0	$\frac{7}{64}$
4	1	1.8	$\frac{15}{16}$	5	1	2.7	$\frac{21}{32}$
3	2	0.2	$\frac{15}{8}$	4	2	0.7	$\frac{105}{64}$
				3	3	0.0	$\frac{35}{16}$

The exact values of  $K$  are given for comparison. For a sample of 5 the critical value is for  $\chi^2$  a shade less than 1.8; but this means a 4:1 sample. For a sample of 6 it lies about midway between a 4:2 and a 5:1 sample, corresponding to  $\chi^2$  about 1.7. We notice, however, that

$K = 0.1$  is not attained by the most extreme samples possible. The interpretation of these small critical values is not that significance can be strongly asserted at them—indeed there is only a probability  $\frac{1}{2}$  of a systematic departure at the critical value anyhow. What they mean is that the outside factor is small, and with the best possible agreement with the null hypothesis there cannot be more than about 2 to 1 support for it. Consequently a smaller value of  $\chi^2$  is needed to reduce  $K$  to 1. The proper conclusion is that where the data are frequencies small samples can tell us little new in any case.

In Tables I and II the values of  $\chi^2$  for given  $K$  increase steadily with  $n$ . I have indicated by italic figures in the upper part of Table I the values that have been calculated, but could not in practice arise in a sampling problem. It is only for a homogeneous sample of 10 that  $K$  can first approach 0.01.

In Tables III and IV the values of  $t^2$  for given  $K$  begin by decreasing as  $\nu$  increases, reach a minimum, and then increase slowly, behaving for large  $\nu$  as  $\chi^2$  does in Tables I and II. The difference is of course due to the allowance for the uncertainty of the standard error, as in the corresponding estimation problems. It is much more important for small  $K$  than for  $K = 1$ .

Table V is intended to test the agreement of a standard deviation with a suggested value.  $K$  is not an even function of  $z$  and therefore it is necessary to tabulate separately for positive and negative  $z$ . It is actually very nearly an even function of  $z/(1 - \frac{1}{2}z)$ , within the range of the table. The asymptotic formula was in satisfactory agreement with the exact formula at  $\nu = 4$ .

It is interesting to compare the results with those based on the customary use of the  $P$  integral. The usual treatment of the problems of Tables I and II would be to draw the line at values of  $\chi^2$  such that they have 5 per cent. or 1 per cent. chances of being exceeded on the null hypothesis. These limits are, for one new parameter, 3.8 and 6.6; for two, 6.0 and 9.2. In Table I,  $K = 1$  lies below the 5 per cent. point up to  $n = 70$ , and passes the 1 per cent. point only about  $n = 1000$ .  $K = 10^{-1/2}$  lies below the 5 per cent. point only for  $n = 5$  and 6, and reaches the 1 per cent. point about  $n = 130$ .

Similarly, in Table II  $K = 1$  lies below the 5 per cent. point up to  $n = 30$ , and passes the 1 per cent. point at  $n = 500$ .  $K = 10^{-1/2}$  never lies below the 5 per cent. point, and reaches the 1 per cent. point about  $n = 40$ .

The 5 per cent. and 1 per cent. points for  $t$  can be taken from the tables given by Fisher, remembering that his  $n$  is my  $\nu$ . The former

drops from  $t^2 = 7.8$  at  $\nu = 4$  to 3.8 for  $\nu$  large; it lies between  $K = 1$  and  $K = 10^{-1/2}$  up to about  $\nu = 50$ , and for larger  $\nu$  below  $K = 1$ . The 1 per cent. point lies between  $K = 10^{-1/2}$  and  $K = 10^{-1}$  up to about  $\nu = 200$ , and below  $K = 1$  for  $\nu = 1000$  and more.

For  $z$  (Table V) the 5 per cent. point and  $K = 1$  are close together both for positive and negative  $z$ . (My negative  $z$  corresponds to Fisher's  $-z$  with  $n_1$  infinite.)  $K = 10^{-1/2}$  agrees fairly well with the 1 per cent. point,  $K = 0.1$  with the 0.1 per cent. point.

In spite of the difference in principle between my tests and those based on the  $P$  integrals, and the omission of the latter to give the increase of the critical values for large  $n$ , dictated essentially by the fact that in testing a small departure found from a large number of observations we are selecting a value out of a long range and should allow for selection, it appears that there is not much difference in the practical recommendations. Users of these tests speak of the 5 per cent. point in much the same way as I should speak of the  $K = 10^{-1/2}$  point, and of the 1 per cent. point as I should speak of the  $K = 10^{-1}$  point; and for moderate numbers of observations the points are not very different. At large numbers of observations there is a difference, since the tests based on the integral would sometimes assert significance at departures that would actually give  $K > 1$ . Thus there may be opposite decisions in such cases. But they will be very rare. We may recall that  $P = 0.01$  means that if  $q$  is true there is a 1 per cent. chance of a larger departure. Hence we can apply Bernoulli's theorem and say that if we assert a genuine departure whenever  $P$  is less than 0.01 we shall expect to be wrong in the long run in 1 per cent. of the cases where  $q$  is true. According to my theory we should expect to make fewer mistakes by taking the limit further out; when  $K = 1$  lies above  $P = 0.01$  there will be a smaller risk of rejecting  $q$  wrongly, partly counter-balanced by a slight increase in the risk of missing a small genuine departure. But in these conditions the probability of a mistake by the use of the 1 per cent. limit for  $P$  is so small anyhow that there is little to be gained by reducing it further. Values between the two limits will be so rare that differences in practice will hardly ever arise. Thus even though the  $P$  tests sometimes theoretically assert  $q'$  when the number of observations is large and my tests support  $q$ , the occasions will be extremely rare.

Actually it may appear that such differences are fairly common; it is known that when the number of observations is very large the estimates of new parameters two to four times the standard error tend to be commoner than would be expected if  $q$  was true, but that these

often or usually do not persist in other similar sets of observations. This, however, is a false contrast, because these discrepancies do not correspond to either the  $q$  or to the  $q'$  of the tests considered in these tables; they represent internal correlation of the errors or non-independence of the chances, and we have not arrived at the hypothesis actually supported by the data until this hypothesis also has been set up and considered. But this leads us to a working rule for saying when such a hypothesis is worth investigation: if an estimate gives  $K > 1$  and  $P < 0.01$ , internal correlation should be suspected and tested, for such a result would not be expected on the hypothesis of independence of the errors in either case. The use of  $P$  by itself involves a danger that discrepancies due to failure of independence will be interpreted as systematic.

TABLE I. *Values of  $\chi^2$  from  $K = \left(\frac{2n}{\pi}\right)^{1/2} \exp(-\frac{1}{2}\chi^2)$*

$n$	$K$				
	1	$10^{-1}$	$10^{-1}$	$10^{-1}$	$10^{-2}$
5	1.2	3.5	5.8	8.1	10.4
6	1.3	3.6	6.0	8.2	10.6
7	1.5	3.8	6.1	8.4	10.7
8	1.6	3.9	6.2	8.5	10.8
9	1.7	4.0	6.3	8.6	10.9
10	1.8	4.2	6.5	8.8	11.1
11	2.0	4.2	6.6	8.9	11.2
12	2.0	4.3	6.6	8.9	11.2
13	2.1	4.4	6.7	9.0	11.3
14	2.2	4.5	6.8	9.1	11.4
15	2.3	4.6	6.9	9.2	11.5
16	2.3	4.6	6.9	9.2	11.5
17	2.4	4.7	7.0	9.3	11.6
18	2.4	4.7	7.0	9.4	11.6
19	2.5	4.8	7.1	9.4	11.7
20	2.5	4.8	7.2	9.4	11.8
30	3.0	5.2	7.6	9.9	12.2
40	3.2	5.5	7.8	10.2	12.4
50	3.5	5.8	8.1	10.4	12.7
60	3.6	5.9	8.2	10.6	12.8
70	3.8	6.1	8.4	10.7	13.0
80	3.9	6.2	8.5	10.8	13.1
90	4.0	6.4	8.7	11.0	13.3
100	4.2	6.4	8.8	11.1	13.4
200	4.8	7.2	9.5	11.8	14.1
500	5.8	8.1	10.4	12.7	15.0
1,000	6.5	8.8	11.1	13.4	15.7
2,000	7.2	9.4	11.8	14.1	16.4
5,000	8.1	10.4	12.7	15.0	17.3
10,000	8.8	11.1	13.4	15.7	18.0
20,000	9.4	11.8	14.1	16.4	18.7
50,000	10.4	12.7	15.0	17.3	19.6
100,000	11.1	13.4	15.7	18.0	20.3

TABLE II.  $\chi^2$  from  $K = \frac{1}{2}\pi n^{1/2}\chi \exp(-\frac{1}{2}\chi^2)$ 

$n$	$K$				
	1	$10^{-1}$	$10^{-1}$	$10^{-1}$	$10^{-2}$
7	4.3	7.1	..	..	..
8	4.5	7.3	..	..	..
9	4.6	7.4	..	..	..
10	4.8	7.5	..	..	..
11	4.9	7.6	..	..	..
12	5.0	7.7	10.3	..	..
13	5.1	7.8	10.4	..	..
14	5.2	7.9	10.4	..	..
15	5.3	8.0	10.5	..	..
16	5.4	8.1	10.6	..	..
17	5.4	8.2	10.7	..	..
18	5.5	8.2	10.8	..	..
19	5.6	8.2	10.8	..	..
20	5.6	8.3	10.9	13.4	15.9
30	6.1	8.8	11.3	13.8	16.3
40	6.5	9.1	11.7	14.2	16.6
50	6.7	9.4	11.9	14.4	16.8
60	6.9	9.5	12.0	14.5	17.0
70	7.0	9.7	12.2	14.7	17.1
80	7.2	9.8	12.3	14.8	17.3
90	7.3	10.0	12.5	15.0	17.4
100	7.5	10.1	12.6	16.1	17.6
200	8.3	10.9	13.4	15.9	18.3
500	9.3	11.9	14.4	16.8	19.3
1,000	10.1	12.6	15.1	17.6	20.0
2,000	10.9	13.4	15.9	18.3	20.7
5,000	11.9	14.4	16.8	19.3	21.7
10,000	12.6	15.1	17.6	20.0	22.4
20,000	13.4	15.9	18.3	20.7	23.2
50,000	14.4	16.8	19.3	21.7	24.1
100,000	15.1	17.6	20.0	22.4	24.8

TABLES OF  $K$ TABLE III.  $t^2$  from  $K = \left(\frac{\pi\nu}{z}\right)^{1/2} \left(1 + \frac{t^2}{\nu}\right)^{-1/2\nu + 1/2}$ 

$\nu$	$K$				
	1	$10^{-1}$	$10^{-2}$	$10^{-3}$	$10^{-4}$
5	3.4	9.9	..	..	..
6	3.4	8.9	17.6	..	..
7	3.4	8.3	15.5	..	..
8	3.5	8.0	14.2	..	..
9	3.5	7.7	13.3	..	..
10	3.6	7.5	12.7	19.2	27.8
11	3.6	7.4	12.2	18.2	25.8
12	3.7	7.3	11.8	17.4	24.2
13	3.7	7.2	11.4	16.8	23.3
14	3.7	7.2	11.2	16.3	22.4
15	3.8	7.1	11.1	15.9	21.5
16	3.8	7.1	11.0	15.4	20.7
17	3.9	7.1	10.9	15.1	20.1
18	3.9	7.0	10.8	14.8	19.6
19	3.9	7.0	10.7	14.6	19.2
20	4.0	7.0	10.6	14.5	18.9
50	4.6	7.4	10.0	12.8	16.0
100	5.2	7.7	10.3	12.8	15.5
200	5.7	8.2	10.7	13.1	15.6
500	6.8	9.1	11.4	13.8	16.2
1,000	7.4	9.7	12.0	14.3	16.6
2,000	8.1	10.4	12.7	15.0	17.3
5,000	9.0	11.3	13.6	15.9	18.2
10,000	9.7	12.0	14.3	16.6	18.9
20,000	10.4	12.7	15.0	17.3	19.6
50,000	11.3	13.6	15.9	18.2	20.5
100,000	12.0	14.3	16.6	18.9	21.2

TABLE IIIA.  $t^2$  from accurate formula 5.2 (33)

$\nu$	$t = 0$		$10^{-1}$	$10^{-2}$	$10^{-3}$	$10^{-4}$
	$K$	$K = 1$				
1	2.3	3.9	30	$1.2 \times 10^4$	$2 \times 10^{13}$	..
2	2.7	3.6	22	102	$10^3$	$10^4$
3	3.0	3.4	12.8	39	120	370
4	3.3	3.4	10.6	26.8	52	118
5	3.5	3.5	9.2	19.4	37	66
6	3.8	3.5	8.5	16.0	29	50
7	4.0	3.5	8.1	15.0	24.2	38
8	4.2	3.6	7.9	13.6	20.6	31
9	4.3	3.8	7.7	13.1	19.5	29.0

TABLE IV.  $t^2$  from  $K = \frac{1}{2}\nu^{1/2}\pi t \left(1 + \frac{t^2}{\nu}\right)^{-1/2}$ 

$\nu$	$K$				
	1	$10^{-1}$	$10^{-1}$	$10^{-1}$	$10^{-1}$
5	7.3	18.4	..	..	..
6	7.0	15.9	..	..	..
7	6.8	14.4	..	..	..
8	6.7	13.1	22.5	35.0	52.2
9	6.7	12.8	20.8	31.3	45.3
10	6.7	12.3	19.4	28.4	40.0
11	6.7	12.0	18.5	26.5	36.7
12	6.7	11.7	17.7	25.0	34.0
13	6.7	11.5	17.2	24.0	32.2
14	6.7	11.3	16.7	23.1	30.6
15	6.7	11.1	16.3	22.3	29.3
16	6.7	11.0	15.9	21.6	28.1
17	6.8	10.9	15.6	21.0	27.2
18	6.8	10.8	15.3	20.5	26.5
19	6.8	10.7	15.1	20.2	25.9
20	6.8	10.7	15.0	19.9	25.3
50	7.3	10.4	13.6	16.9	20.3
100	7.9	10.8	13.6	16.4	19.3
200	8.5	11.2	13.9	16.5	19.2
500	9.4	12.0	14.6	17.2	19.7
1,000	10.2	12.8	15.2	17.7	20.2
2,000	10.9	13.4	15.9	18.3	20.8
5,000	11.9	14.4	16.8	19.3	21.7
10,000	12.7	15.1	17.6	20.0	22.4
20,000	13.4	15.9	18.3	20.8	23.2
50,000	14.4	16.9	19.3	21.7	24.1
100,000	15.1	17.6	20.0	22.4	24.8

TABLE IV A.  $t^2$  from accurate formula 6.21 (42)

$\nu$	$t = 0$	$K = 1$	$10^{-1}$	$10^{-1}$	$10^{-1}$	$10^{-2}$
	$K$					
1	2.7	9.1	1,500	$10^{10}$	..	..
2	3.0	6.8	48	380	3,300	32,000
3	3.3	6.5	24.5	79	251	790
4	3.5	6.2	18.2	43.6	100	216
5	3.8	6.1	15.7	33.6	70	138
6	4.0	6.0	13.9	26.6	49	85
7	4.2	5.9	12.8	22.2	36	55
8	4.3	5.9	12.3	20.7	32.6	49.1



TABLE V.  $z$  from 5.43 (11) and (14)

$\nu$	$z=0$		$10^{-1}$	$10^{-1}$	$10^{-1}$	$10^{-1}$	$K=1$	$10^{-1}$	$10^{-1}$	$10^{-1}$	$10^{-1}$
	$K$	$K=1$									
1	1.8	+0.77	+1.04	+1.20	+1.31	+1.40	-1.4	-8.5	?	?	?
2	2.2	+0.56	+0.76	+0.94	+1.04	+1.12	-1.13	-2.2	-3.2	-4.4	-5.5
3	2.5	+0.47	+0.70	+0.78	+0.86	+0.94	-0.99	-1.68	-2.30	-2.88	-3.46
4	2.8	+0.45	+0.62	+0.72	+0.82	+0.89	-0.70	-1.17	-1.60	-2.01	-2.42
5	3.1	+0.43	+0.57	+0.67	+0.75	+0.82	-0.65	-1.04	-1.38	-1.71	-2.03
6	3.4	+0.41	+0.54	+0.63	+0.70	+0.77	-0.61	-0.94	-1.21	-1.47	-1.72
7	3.6	+0.39	+0.51	+0.60	+0.65	+0.73	-0.57	-0.85	-1.08	-1.30	-1.51
8	3.9	+0.37	+0.49	+0.57	+0.63	+0.69	-0.52	-0.77	-0.98	-1.18	-1.36
9	4.1	+0.36	+0.46	+0.54	+0.60	+0.66	-0.49	-0.71	-0.90	-1.08	-1.25
10	4.2	+0.34	+0.44	+0.52	+0.58	+0.63	-0.47	-0.67	-0.85	-1.01	-1.18
12	4.6	+0.32	+0.42	+0.49	+0.54	+0.59	-0.43	-0.60	-0.75	-0.89	-1.02
14	4.9	+0.31	+0.39	+0.46	+0.51	+0.55	-0.40	-0.55	-0.68	-0.81	-0.92
16	5.2	+0.30	+0.37	+0.43	+0.48	+0.52	-0.38	-0.51	-0.63	-0.74	-0.85
18	5.5	+0.29	+0.36	+0.41	+0.46	+0.50	-0.36	-0.48	-0.59	-0.71	-0.78
20	5.8	+0.27	+0.34	+0.40	+0.44	+0.48	-0.34	-0.45	-0.55	-0.68	-0.73
50	9.0	+0.20	+0.24	+0.27	+0.30	+0.33	-0.22	-0.29	-0.34	-0.38	-0.43

## NOTE ON THE CONSISTENCY OF THE PRODUCT RULE

WE assume weaker forms of Axioms 1, 2, 3, 4, 5, 6, namely that they hold on a sufficiently general datum  $H$ . Any actual datum is supposed to contain  $H$ . We use Conventions 1 and 2 on  $H$  and assume that Convention 3 is applicable on  $H$ . Then Theorems 1, 2, 3, 4, 5, 6, 7, 8 follow if the datum is  $H$ .

Now if  $p$  is an additional datum such that  $P(p|H) \neq 0$ , and  $q_i$  are a set of propositions, exhaustive on  $H$ , whose disjunction is  $Q$ , we assume

$$P(q_i|pH) = \frac{P(pq_i|H)}{P(p|H)}. \quad (1)$$

This provides the first means, in this presentation, of calculating probabilities when  $H$  is not the only datum. Convention 1 on  $pH$  becomes a rule for the ordering of probabilities in terms of their numerical assessments instead of conversely.

Since the  $P(pq_i|H)$  satisfy Ax. 1 and  $P(p|H)$  is independent of  $q_i$ , it follows that the  $P(q_i|pH)$  satisfy Ax. 1. Similarly they satisfy Ax. 2, 4, Conv. 2 (since if  $q_i, q_j$  are exclusive on  $H$  they are also exclusive on  $pH$ ; and if  $q_i, q_j$  are exclusive on  $pH$ ,  $pq_i, pq_j$  are exclusive on  $H$ ), and Ax. 5.

For Ax. 6, we have, if  $pq_i$  entails  $r_k$ ,

$$P(q_i r_k | pH) = \frac{P(pq_i r_k | H)}{P(p | H)} = \frac{P(pq_i | H)}{P(p | H)} = P(q_i | pH),$$

using Ax. 6 on data  $H$ ; hence Ax. 6 holds on data  $pH$ .

Next, if  $pH$  entails  $q_i$ ,  $P(pq_i|H) = P(p|H)$  by Ax. 6, and therefore  $P(q_i|pH) = 1$ . Conv. 3 becomes a theorem, and the first part of Ax. 3 follows. If  $pH$  entails  $\sim q_i$ ,  $pq_i$  is impossible given  $H$  and therefore  $P(pq_i|H) = 0$ ,  $P(q_i|pH) = 0$ ; hence we have the second part of Ax. 3.

For Ax. 7, consider two sets of propositions each exhaustive on  $H$ , say  $q_i, r_k$ ; then Ax. 7 will read

$$P(q_i r_k | pH) = P(q_i | pH) P(r_k | q_i pH) / P(p | q_i pH). \quad (2)$$

By (1) this is equivalent to

$$\frac{P(pq_i r_k | H)}{P(p | H)} = \frac{P(pq_i | H)}{P(p | H)} \frac{P(pq_i r_k | H)}{P(pq_i | H)} \bigg/ \frac{P(pq_i p | H)}{P(pq_i | H)},$$

which is an identity. Hence Ax. 7 follows.

In this presentation we assume no properties of probabilities on data other than  $H$ , except that they can be calculated by (1), and this is possible if the axioms are satisfied by probabilities on  $H$ . Hence if pure

mathematics and the axioms on  $H$  are consistent, the axioms remain consistent when applied to probabilities on data including  $H$ .

An apparent difficulty about this argument as a general proof of consistency is that if  $H$  is the general principles of the theory and  $p$  a special proposition, we may not be able to use Conv. 3 on data  $H$ . This can be met in two ways. We have seen that the principle of inverse probability is consistent if the product rule is consistent for likelihoods, and therefore it is enough if  $H$  in the argument includes a law such that we can use Conv. 3; but this is always true for likelihoods. The other way is to notice that if  $H$ , for instance, expresses ignorance of a standard error, we may arbitrarily impose bounds on the possible values so that  $0 < \sigma_1 \leq \sigma \leq \sigma_2 < \infty$  and use Conv. 3; and our results will be consistent as limits of the results when  $\sigma_1 \rightarrow 0$ ,  $\sigma_2 \rightarrow \infty$ , and infinite integrals are interpreted in this way in any case. This way of looking at the matter may be preferred. For if  $H'$  is such that we can use Conv. 3 on it, and  $H$  differs from  $H'$  only by including the statement that a standard error is unknown, then all non-zero probabilities on  $H'$  are replaced by infinite ones on  $H$ ; a statement that we do not know a standard error is apparently accompanied by an instruction to forget for a time everything that we ever knew.

## NOTE ON THE INFINITE REGRESS ARGUMENT

THE customary procedure in a mathematical system is to state a set of definitions and postulates and to examine what consequences follow from them. It is often said that all concepts should be defined and all postulates should be proved. It is worth while to point out that to admit this would invalidate any argument. Suppose that a system starts from concepts  $A_1, A_2, \dots$  and postulates  $p_1, p_2, \dots$ , and that we are required to define  $A_1$ . We may be able (1) to define it in terms of  $A_2, A_3, \dots$ , or (2) to define it in terms of a concept  $A'_1$  not included in  $A_2, A_3, \dots$ . If (1) is possible the number of fundamental concepts is reduced; but repetition of the process for  $A_2$  reproduces the same situation. Suppose then that we find a set  $B_1, B_2, \dots$ , none of which can be defined in terms of the others, and are asked to define  $B_1$ . The definition must be in terms of a further concept  $C_1$ , which would therefore have to be defined in terms of  $D_1$ , and so on for ever. Hence we can never define all the concepts of a system.

Similarly to prove  $p_1$  would require a proof from  $p_2, p_3, \dots$  or the introduction of a new postulate, and again we should always find at some stage that the proof of a postulate requires the introduction of a new one.

An argument, the application of which would always lead to the introduction of a new definition or postulate not within the system, is said to involve an *infinite regress*. Several arguments in the text are designed to avoid infinite regresses (pp. 112, 116, 375), but the principle is not stated in general terms.

A famous example is Lewis Carroll's† 'What the Tortoise said to Achilles'. The propositions  $p$  and  $p$  implies  $q$  imply  $q$ . But if we accept  $p$  and  $p$  implies  $q$  we cannot symbolize a proof that we can assert  $q$  by itself. If we try we find ourselves in an infinite regress. The use of 'therefore' can be stated, understood, and acted on only verbally, not symbolically.

† *Complete Works*, 1225-30; *Mind*, 4, 1895, 278-80.

## INDEX

- Abbreviations  $x$ ,  $dx$ , 120.  
 Accidents, factory, 69, 295.  
 Accuracy, useful degree of, 125, 397.  
 Adams, J. C., 389.  
 Addition rule, 19, 30, 33.  
 Agreement, too close, 281.  
 Agricultural experiments, 127, 214, 361.  
 Aitken, John, 60.  
 Alternative hypothesis, 220.  
 Amoeba, 6.  
 Ancillary statistics, 182.  
 Applicability, 8, 9, 11.  
 Approximations, 50, 140, 168, 170, 251.  
*A priori*, 8, 29.  
 Argon, 260.  
 Arithmetic mean, 84, 92, 107, 189.  
 Assent, universal, 14, 46.  
 Atmospheric tide, 307.  
 Average residual, 188.  
  
 Barnes, E. W., 395.  
 Bartlett, M. S., 41, 53, 147.  
 Bateman, H., 59.  
 Bayes, T., 29, 30, 34, 42, 102, 107, 109, 374.  
 Behaviourism, 45, 379.  
 Belief, *see* Confidence.  
 Bellamy, Miss E. F., 324.  
 Benefit, expectation of, 30.  
 Bernoulli, Daniel, 32.  
 Bernoulli, James, 52.  
 Bias, 143, 177, 231.  
 Binomial law, 50, 56.  
 Binomial, negative, 68, 77.  
 Black, A. N., 174.  
 Boltzmann, 28, 369.  
 Bond, W. N., 288.  
 Bortkiewicz, L. von, 59.  
 Boys, C. V., 280.  
 Broad, C. D., 5, 26, 111, 112, 115, 372.  
 Brown, E. W., 362.  
 Brunt, D., 211, 269.  
 Bullard, E. C., 84, 129, 137.  
 Bullen, K. E., 175.  
 Burnside, W., 345.  
  
 Campbell, N. R., 6, 14, 41.  
 Cantelli, F. P., 55.  
 Carnap, R., 20.  
 Carroll, Lewis, 45, 220, 305, 407.  
 Cauchy rule, 78, 81, 170, 189, 244.  
 Causality, 12, 108.  
 Caution, 273, 287, 381.  
 Central limit theorem, 79.  
 Certainty, approach to, 38, 336.  
     on the data, 17.  
 Chance, 41, 50, 229.  
     continuous distribution, 301.  
     games of, 32, 47.  
 Chapman, S., 307.  
 Characteristic function, 73.  
 Chauvenet, 188, 291, 357.  
  
 Checking, 139.  
 Chosterton, G. K., 395.  
 Combination of estimates, 175.  
     of tests, 305.  
 Common sense, 1.  
 Comparison of chances, 235.  
 Comrie, L. J., 62.  
 Confidence, reasonable degree of, 15.  
 Conjunction, 18.  
 Consistency, 8, 19, 35, 36, 159, 166, 170, 251, 405.  
 Continental drift, 48.  
 Contingency, 211, 232.  
     diagonal elements, 332.  
 Continuity, 21, 24.  
 Continuous variation, 227.  
 Conventions, 20, 30.  
 Correlation, 71, 152, 263.  
     correction of, 202.  
     internal, 271, 287, 289, 400.  
     intraclass, 72, 198, 268, 276, 314.  
     partial, 328.  
     rank, 204, 268.  
     serial, 170, 227, 328.  
 Cournot, A., 374.  
 Cramér, H., 80, 343.  
 Critical realism and idealism, 46.  
 Curvature of universe, 304.  
  
 Damoiseau, 389.  
 Darwin, Sir G. H., 390.  
 Data, need to state, 15, 27, 350, 377.  
 Deduction, 1, 3, 17.  
     as approximation, 336.  
 Definitions, 379.  
 Degrees of freedom, 89, 128.  
 Delauney, 389.  
 De Moivre, A., 52, 342.  
 Density, probability, 24.  
 Design of experiments, 97, 214, 361.  
 Determinism, 11.  
 Deviation, standard, 92, 128.  
 Diananda, P. H., 60, 169.  
 Dice, 50, 231, 306, 314.  
 Digamma function, 187.  
 Dingle, H., 14, 384.  
 Dip, magnetic, 84.  
 Dirichlet integrals, 87, 116.  
 Disjunction, 18.  
     separation of, 41.  
 Dodgson, C. L., *see* Carroll, Lewis.  
 Dust counter, 60, 241.  
  
 Earthquakes, aftershocks, 325, 334.  
     determination of epicentres, 136.  
     identity of epicentres, 322.  
     law of error, 190.  
     periodicity, 324.  
     travel times:  
         *P*, 175, 202, 273, 299.  
         *S* and *SKS*, 265.

- Economy of thought, 4.  
 of postulates, 9, 37, 46, 102, 345, 384.  
 Eddington, Sir A. S., 6, 193, 283, 379, 386.  
 Edgeworth, F. Y., 108.  
 Efficiency, 145, 179.  
 Einstein, A., 362, 386.  
 Ellis, R. L., 345, 374.  
 Emmett, W. G., 365.  
 Ensemble, 11, 341.  
 Entailment, 17, 48.  
 Epistemology, 1, 12, 13.  
 Equations of condition, 133.  
 normal, 133.  
 Ergodic theory, 371.  
 Errors, 12, 13.  
 accidental, 270.  
 composition of, 74, 79.  
 independence of, 286; *see also* correlation, internal.  
 normal law, 60, 287.  
 probable, 62, 124.  
 standard, 62.  
 systematic, 270, 273.  
 unknown law of, 187.  
*See also* Pearson types.  
 Estimation, 99.  
 relation to significance, 359.  
 Euclid, 8, 40.  
 Exclusive, 18.  
 Exhaustive, 18.  
 Expectation, mathematical, 31, 43, 177, 364.  
 moral, 31.  
 of benefit, 31, 43.  
 Explanation, 303, 390.  
 Factorial function, 51, 239.  
 Factory accidents, 69, 295.  
 Fiducial argument, 352.  
 Fieller, E. C., 55.  
 Fisher, R. A., 11, 29, 63, 88, 92, 96, 124, 127, 152, 169, 179, 181, 184, 189, 197, 209, 214, 229, 238, 282, 306, 341, 352, 356, 364.  
 Fowler, Sir R. H., 370.  
 Franks, W. S., 211.  
 Fréchet, M., 371.  
 Freedom, degrees of, 89, 128.  
 Freeman, R. A., 100, 392.  
 Frequency definitions, 11, 34, 342, 372.  
 Freud, S., 239.  
 Functions, new, 295.  
 old, 299.  
 Galton, Sir F., 72.  
 Gases, kinetic theory, 28, 369.  
 Gauss, C. F., 14, 62, 84, 103, 133, 190.  
 Geiger, H., 59.  
 Gender, 239.  
 Generalization (empirical propositions), 1, 3.  
 (logical propositions), 7, 26.  
 George, W. H., 12.  
 Gibbs, Willard, 11, 341, 369.  
 Glaisher, 390.  
 Gödel, K., 35, 56.  
 Gosset, W. L., *see* 'Student'.  
 Grades, 206, 210.  
 Gravitation, law of, 362.  
 constant of, 280.  
 Gravity, 129, 137, 198.  
 Greenwood, M., 69.  
 Grouping, 136, 184, 193, 326.  
*H* (definition), 48.  
 Haldane, J. B. S., 107, 120, 162.  
 Heisenberg, H., 14.  
 Heyl, P. R., 280.  
 Hilbert, 10.  
 Hill, G. W., 362.  
 Horse, kicks by, 59, 71, 295.  
 Hosiasson, Miss J., 342.  
 Hulme, H. R., 288.  
*I<sub>m</sub>* defined, 158.  
 Idealism, 49, 393.  
 Ignorance, 34, 101, 220, 222, 353.  
 Implication, 17, 48.  
 Impossibility, 17.  
 Induction, 1, 8.  
 Infinite population, 11, 341, 345.  
 Infinite regress, 112, 116, 375, 407.  
 Inoculation, 239, 312.  
 Insufficient reason, 34.  
 Integer, unknown, 213.  
 Internal correlation, 271, 287, 289, 400.  
 Intuition, 15.  
 Invariance, 104, 158, 170, 248.  
 Inverse probability, 29, 35, 372.  
 Irrelevance, 28, 41, 42, 160, 163.  
*J* defined, 158.  
 Jeans, Sir J. H., 370.  
 Johnson, W. E., 19, 26, 118, 372.  
 Joint assertion, 18.  
 Jolly, H. L. P., 137.  
 Jones, Sir H. Spencor, 278.  
 Jourdain, P. E. B., 38.  
*K* defined, 221.  
 tables, 306.  
 Kapteyn, 269.  
 Kendall, M. G., 49, 81, 88, 96, 210, 239, 281, 326, 372.  
 Keynes, Lord, 26, 59, 147.  
 Knott, C. G., 326.  
 Knowledge, vague, 107, 121, 152, 219, 225, 306, 377.  
 Lagrange, J. L., 376.  
 Lange, J., 238, 355.  
 Language, 19, 20, 32, 45, 372, 378, 391.  
 Laplace, P. S., 14, 23, 29, 31, 34, 62, 102, 107, 133, 374, 389.  
 rule of succession, 110.  
 Latin square, 215.  
 Law of large numbers, 52.  
 Law, scientific, 3, 13, 99, 113, 220, 336, 349.

- Least squares, 129.  
   approximations, 140, 173.  
 Le Verrier, U. J. J., 389.  
 Likelihood, 29, 47, 99.  
   maximum, 168, 170, 189.  
 Limit of sampling ratio, 53, 341, 345.  
 Littlewood, J. E., 56, 76.  
 Location parameter, 63.  
 Logical product, 18, 25.  
   quotient, 25.  
   sum, 18, 25.  
 Lüders, R., 70.  
  
 McColl, H., 26.  
 Materialism, 394.  
 Mathematics, pure, 2, 10, 37.  
   applied, 2, 3, 12.  
 Maximum likelihood, 168, 170, 189.  
   relation to invariance theory, 169.  
 Maxwell, J. C., 1, 369.  
 Mean square contingency, 212.  
   deviation, 92.  
 Measures, significance tests, 242, 251, 315.  
 Median law, 76, 78, 188.  
 Median, use of, 187, 293.  
   of general law, 148.  
 Mendelism, 108, 282, 311, 360.  
 Mercury, perihelion of, 387.  
 Metaphysics, 394.  
 Method and material, 7, 9, 10, 388.  
 Milne, E. A., 6.  
 Milne-Thomson, L. M., 62.  
 Mind, human, 5, 9, 37, 107, 377, 392.  
 Mises, R., 341, 345.  
 Moments, 73, 74, 76, 183.  
 Moon, secular acceleration of, 389.  
 Moore, G. E., 17.  
 Muirhead, J. H., 14.  
 Multinomial law, 57, 90.  
 Multiple sampling, 57.  
 Multiplicative axiom, 10, 56.  
  
 Naïve realism and idealism, 46, 383.  
 Negative binomial, 68, 77, 293.  
 Newall, H. F., 387.  
 Newbold, Miss E. M., 295.  
 Newman, M. H. A., 213.  
 Newton, 40, 340, 362.  
 Neyman, J., 172, 177, 341, 343, 366.  
 Nitrogen, density of, 260.  
 Normal equations, 133.  
 Normal law, derivation, 60, 79.  
   departure from, 190.  
   estimation problem, 120.  
   moments, 78.  
   reproductive property, 79.  
   significance tests for parameters, 242, 251.  
   test of, 287.  
 Nutation, 278.  
 Null hypothesis, 229.  
 Numbers, introduction of, 19.  
  
 Objectivity, 11, 376.  
 Observations, rejection of, 188.  
  
 Ockham, 315, 385.  
 Offord, A. C., 150.  
  
*P* integral, 355, 398.  
 Pairman, Miss E., 187.  
 Paneth, F. A., 263.  
 Parallax, negative, 142, 204.  
   stellar, 300.  
 Parameters, number admissible, 100, 315.  
   location and scale, 63.  
   old and new, 222.  
   orthogonal, 184, 223.  
   suggested values, 109.  
 Pearson, E. S., 177, 219, 271, 366.  
 Pearson, Karl, 7, 45, 62, 72, 88, 108, 115, 125, 172, 183, 204, 231, 270, 288, 310, 354, 374.  
 Pearson types, 64, 185.  
 Peirce, C. S., 188.  
 Periodicity, 315.  
 Perks, W., 170.  
 Personal equation, 270.  
 Petersburg problem, 32.  
 Physicists, old-fashioned, 244, 274.  
 Plana, 389.  
 Poisson law, 58, 68, 77, 119, 237, 240, 293.  
 Ponce, John, 315.  
 Pontécoulant, 389.  
 Postulates, economy of, 9, 37, 46, 102.  
 Precision constant, 62.  
 Prediction, 1, 4, 13, 14, 40.  
*Principia Mathematica*, 6, 8, 10, 18, 25, 48, 391.  
 Probability, 15.  
   aim of theory, 8.  
   density, 24.  
   posterior, 29.  
   prior, 29, 34.  
   invariance rules:  
     estimation, 158.  
     significance, 248.  
   logarithmic rule, 102, 104, 119.  
   of laws, 100.  
   revision of, 310.  
   truncation of, 142, 197, 203.  
   uniform rule, 102.  
 Probable error, 62, 124.  
 Product rule, 25.  
   consistency of, 35, 36, 405.  
   incorrect form of, 27.  
 Psychoanalysis, 239.  
 Psychology, 37, 38.  
  
 Quantum theory, 100, 382, 387.  
 Questions, statement of, 91, 108.  
 Quinney, H., 378.  
 Quotient, logical, 26.  
  
 Radioactivity, 59, 71, 241.  
 Ramsay, F. P., 10, 26, 31, 372.  
 Randomization, 214, 272, 297.  
 Randomness, 49.  
   rule of procedure, 315, 385.  
 Rank correlation, 204.  
 Rayleigh, Lord, 260.

- Reading of scale, 146.  
 Realism, 44, 393.  
 Reality, 338.  
 Rectangular law, 66, 85, 143, 184.  
 Reduction, uniform, 192.  
 Regression, 73.  
 Rejection of observations, 188, 280, 287.  
   of unobservables, 383, 387.  
 Relativity, 39, 48, 385.  
 Religion, 394.  
 Re-scaling of law, 145.  
 Residuals, 133, 188.  
 Rounding-off errors, 85, 146, 195.  
 Russell, Bertrand, 5, 46, 380; *see also*  
   *Principia Mathematica*.  
 Rutherford, Lord, 59.  
  
 Sadler, D. H., 362.  
 Samples, comparison of, 235.  
 Sampling, simple, 49, 56, 109.  
   multiple, 57, 117.  
   with replacement, 50.  
 Scale parameter, 63.  
 Scale, reading of, 146.  
 Schuster, Sir A., 227, 326.  
 Scrase, F. J., 60.  
 Seidel, 140, 173.  
 Selection, allowance for, 226.  
 Sheppard, W. F., 62, 195.  
 Significance, 100, 220.  
   approximate form, 251.  
   combination of tests, 305.  
   complications, 222.  
   invariance, 248.  
 Simplicity, 4, 100, 103, 113, 222, 391.  
 Smithies, F., 376.  
 Smoothing, 198.  
 Solipsism, 44, 379, 393.  
 Southwell, R. V., 174.  
 Spearman, C., 204.  
 Standard deviation, 92, 128, 133.  
 Standard error, 62.  
   errors, agreement of, 242.  
 Stars, colour and spectral type, 210.  
 Statistical mechanics, 28, 369.  
 Statistics, sufficient, 92.  
   ancillary, 182.  
   efficiency of, 179.  
   unbiased, 177.  
 Stebbing, L. S., 14.  
 Stevens, W. L., 333.  
 Stieltjes integral, 73.  
 Stirling's formula, 51.  
 Storer, W. O., 127.  
 Struggle for existence, 6.  
 'Student', 94, 122, 205, 219, 271, 350,  
   364.  
 Succession, rule of, 110.  
  
 Suggested values, 108.  
 Survey, Ordnance, 175.  
  
 $t$  rule, 95, 122, 124, 128.  
   significance, 242, 316, 319, 402, 403.  
 Taylor, Sir G. I., 332, 378, 382, 390.  
 Telepathy, 333.  
 Teodorescu, 258.  
 Theory, 390.  
 Thorburn, W. M., 315.  
 Tidal friction, 390.  
 Tires, strength of, 258.  
 Titchmarsh, E. C., 76.  
 Triangular distribution, 85.  
 True value, 62.  
   significance tests, 242.  
 Turbulence, 332.  
 Turner, H. H., 227.  
 Twins, 238, 312.  
  
 Uncertainty principle, 13.  
 Undistributed middle, 2, 39, 381.  
 Unforeseen alternative, 39, 381.  
 Uniform reduction, 192.  
 Uniformity of Nature, 5, 11.  
 Universal assent, 14.  
 Unobservables, 383, 387.  
  
 Venn limit, 11, 341, 345.  
 Venus, node of, 362.  
  
 Walker, Sir G. T., 229.  
 Watson, G. N., 83.  
 Weber and Fechner, 32.  
 Weight, 124, 136.  
 Weldon, W. F. R., 231, 314, 340.  
 Whipple, F. J. W., 109, 328.  
 Whitehead, A. N., *see Principia Mathe-*  
   *matica*.  
 Whittaker, Sir E. T., and Robinson, G.,  
   80, 84, 202.  
 Wish-fulfilment, 16, 392.  
 Wrinch, D., 26, 53, 100, 112.  
  
 Yamaguti, S., 327.  
 Yates, F., 209, 214, 219, 281, 357.  
 Yule, G. Udny, 49, 69, 88, 96, 208, 239,  
   326, 356.  
  
 $z$  rule, 95, 125.  
   modification of, 97.  
   significance, 255, 257, 404.  
  
 $\theta$ , definition, 48.  
 $\chi^2$ , 85, 87, 363.  
 $\chi^2$  with estimated standard errors, 97.  
   too small, 281.  
 $\chi'^2$ , 170.



PRINTED IN  
GREAT BRITAIN  
AT THE  
UNIVERSITY PRESS  
OXFORD  
BY  
CHARLES BATEY  
PRINTER  
TO THE  
UNIVERSITY

